



RUDN
university



DISTRIBUTED COMPUTER AND COMMUNICATIONS NETWORKS:
CONTROL, COMPUTATION, COMMUNICATIONS (DCCN-2022)

Russian Academy of Sciences (RAS)
V.A. Trapeznikov Institute of Control Sciences of RAS (ICS RAS)
Peoples' Friendship University of Russia (RUDN University)
Institute of Information and Communication Technologies
of Bulgarian Academy of Sciences (Sofia, Bulgaria)
National Research Tomsk State University (NRTSU)
Research and development company
"Information and networking technologies"

DISTRIBUTED COMPUTER
AND COMMUNICATION NETWORKS:
CONTROL, COMPUTATION,
COMMUNICATIONS
(DCCN-2022)



PROCEEDINGS
OF THE XXV INTERNATIONAL SCIENTIFIC CONFERENCE
Russia, Moscow, September 26–30, 2022



Moscow
Peoples' Friendship University of Russia
2022

Российская академия наук (РАН)
Институт проблем управления им. В.А. Трапезникова
Российской академии наук (ИПУ РАН)
Российский университет дружбы народов (РУДН)
Институт информационных и телекоммуникационных технологий
Болгарской академии наук (София, Болгария)
Национальный исследовательский Томский государственный университет (НИ ТГУ)
Научно-производственное объединение
«Информационные и сетевые технологии» («ИНСЕТ»)

РАСПРЕДЕЛЕННЫЕ КОМПЬЮТЕРНЫЕ И ТЕЛЕКОММУНИКАЦИОННЫЕ СЕТИ: УПРАВЛЕНИЕ, ВЫЧИСЛЕНИЕ, СВЯЗЬ (DCCN-2022)



DCCN
2022

Материалы
XXV Международной научной конференции

Россия, Москва, 26–30 сентября 2022 г.

Под общей редакцией
д.т.н. *В.М. Вишневского* и д.т.н. *К.Е. Самуилова*

Москва
Российский университет дружбы народов
2022

Russian Academy of Sciences (RAS)
V.A. Trapeznikov Institute of Control Sciences of RAS (ICS RAS)
Peoples' Friendship University of Russia (RUDN University)
Institute of Information and Communication Technologies
of Bulgarian Academy of Sciences (Sofia, Bulgaria)
National Research Tomsk State University (NR TSU)
Research and development company
"Information and networking technologies"

**DISTRIBUTED COMPUTER
AND COMMUNICATION NETWORKS:
CONTROL, COMPUTATION,
COMMUNICATIONS
(DCCN-2022)**



**Proceedings
of the XXV International Scientific Conference**

Russia, Moscow, September 26–30, 2022

Under the general editorship
of D.Sc. *V.M. Vishnevskiy* and D.Sc. *K.E. Samouylov*

Moscow
Peoples' Friendship University of Russia
2022

Под общей редакцией
д.т.н. В.М. Вишневского и д.т.н. К.Е. Самуйлова

P24 **Распределенные компьютерные и телекоммуникационные сети : управление, вычисление, связь (DCCN-2022) = Distributed computer and communication networks : control, computation, communications (DCCN-2022) :** материалы XXV Международной научной конференции. Россия, Москва, 26–30 сентября 2022 г. / под общ. ред. В. М. Вишневского и К.Е. Самуйлова. – Москва : РУДН, 2022. – 439 с. : ил.

В научном электронном издании представлены материалы XXV Международной научной конференции «Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь» по следующим направлениям:

- Алгоритмы и протоколы телекоммуникационных сетей;
- Управление в компьютерных и инфокоммуникационных системах;
- Анализ производительности, оценка QoS / QoE и эффективность сетей;
- Аналитическое и имитационное моделирование коммуникационных систем последующих поколений;
- Эволюция беспроводных сетей в направлении 5G;
- Технологии сантиметрового и миллиметрового диапазона радиоволн;
- RFID-технологии и их приложения;
- Интернет вещей и туманные вычисления;
- Системы облачного вычисления, распределенные и параллельные системы;
- Анализ больших данных;
- Вероятностные и статистические модели в информационных системах;
- Теория массового обслуживания, теория надежности и их приложения;
- Высотные беспилотные платформы и летательные аппараты: управление, передача данных, приложения.

В материалах научной конференции DCCN-2022 обсуждены перспективы развития и сотрудничества в этой сфере.

Издание предназначено для научных работников и специалистов в области управления крупномасштабными системами.

Текст воспроизводится в том виде, в котором представлен авторами.

Конференция организована при поддержке Программы стратегического академического лидерства РУДН.

Содержание / Contents

1. Аминев Д.А., Богданова Е.Г., Козырев Д.В.	
ТРЕБОВАНИЯ СЦЕНАРИЕВ URLLC, mMTC, eMMB К УРОВНЯМ L0, L1 ТРАНСПОРТНОЙ СЕТИ IMT-2020/5G	1
2. Markovich N.M., Ryzhov M.S.	
ESTIMATION OF THE TAIL INDEX OF PAGERANKS IN RANDOM GRAPHS	15
3. Ткачев О.А.	
ОПРЕДЕЛЕНИЕ СРЕДНЕГО ВРЕМЕНИ РАБОТЫ ДО ОТКАЗА БЕСПРОВОДНОЙ СЕНСОРНОЙ СЕТИ	22
4. Rykov V., Ivanova N.	
RELIABILITY OF A LOAD-SHARING K-OUT-OF-N SYSTEM UNDER DECREASING OF COMPONENTS RESIDUAL LIFETIME	28
5. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V.	
THE TIMELINESS OF THE RESERVED SERVICE IN THE CLUSTER WITH THE REGULATION OF THE TIME OF DESTRUCTION OF OVERDUE REQUESTS IN THE NODE QUEUES	34
6. Markovich N.M., Ryzhov M.S.	
CLUSTERS OF EXCEEDANCES FOR EVOLVING RANDOM GRAPHS.....	44
7. Nguyen V.T., Pashchenko F.F., Bui T.A., Le D.T.	
IMPROVEMENT OF CNN-BASED MODEL FOR OBJECT CLASSIFICATION IN AERO PHOTOGRAPHS	51
8. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V.	
MULTIPATH REDUNDANT TRANSMISSION WITH TRAFFIC HETEROGENEITY IN TERMS OF THE CRITICALITY OF NETWORK DELAYS	65
9. Grebeshkov A. Y.	
IIOT INFORMATION PROCESSING MODEL FOR TRANSFER LEARNING WITH DATA QUALITY MANAGEMENT	77
10. Михайлов К.И., Абрамов А.Г.	
ТЕОРИЯ И ПРАКТИКА ОПРЕДЕЛЕНИЯ УРОВНЯ КРИТИЧНОСТИ ИНЦИДЕНТОВ В ЦИФРОВЫХ ИНФРАСТРУКТУРАХ	83
11. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V.	
RELIABILITY OF A REDUNDANT COMPUTER SYSTEM, TAKING INTO ACCOUNT THE FEATURES OF INFORMATION RECOVERY	89
12. Vorobeychikov S.E., Pupkov A.V.	
NON-ASYMPTOTIC CONFIDENCE ESTIMATION OF THE AUTOREGRESSIVE PARAMETER IN ARMA(1,Q) MODEL.....	101
13. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V.	
CLUSTER WITH FUNCTIONAL HETEROGENEITY OF NODES WITH REQUESTS OF DIFFERENT CRITICALITY TO DELAYS	107
14. Bui T.A., Pashchenko F.F., Tran D.H., Nguyen V.T., Pham T.N.	
NEURO-FUZZY MODEL BASED ON MULTIDIMENSIONAL MEMBERSHIP FUNCTION.....	119

15. Sztrik J., Tóth Á., Pint’er Á., B’acs Z. PERFORMANCE ANALYSIS OF A FINITE-SOURCE RETRIAL QUEUEING SYSTEM WITH TWO-WAY COMMUNICATION, CATASTROPHIC BREAKDOWN AND IMPATIENT CUSTOMERS USING SIMULATION.....	129
16. Song J., Namiot D. ON MODEL INVERSION ATTACKS	135
17. А.З. Меликов А.З., Мирзоев Р.Р., Наир С.С. МЕТОД РАСЧЕТА ХАРАКТЕРИСТИК СИСТЕМЫ ОБСЛУЖИВАНИЯ С ГИБРИДНОЙ ПОЛИТИКОЙ ПОПОЛНЕНИЯ ЗАПАСОВ ОТ ДВУХ ИСТОЧНИКОВ.....	143
18. Namiot D., Ilyushin E. ON MONITORING OF MACHINE LEARNING MODELS.....	150
19. Hilquias V.C.C., Zaryadov I.S., Milovanova T. A. QUEUEING SYSTEM WITH THRESHOLD-BASED GENERAL RENOVATION MECHANISM.....	158
20. Kochueva O.N. FEATURE SELECTION FOR A FUZZY CLASSIFICATION MODEL BASED ON A GENETIC ALGORITHM	168
21. Плаксин Д.А., Фёдорова Е.А., Лизюра О.Д., Шашев Д.В., Моисеева С.П. МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПЕРЕДАЧИ ДАННЫХ В СЕТИ FANET В ВИДЕ RQ-СИСТЕМ	176
22. Назаров А.А., Фунг-Дук Т., Пауль С.В., Лизюра О.Д. АСИМПТОТИЧЕСКИ-ДИФФУЗИОННЫЙ АНАЛИЗ RQ-СИСТЕМЫ MMRP/M/1 С РАЗНОТИПНЫМИ ВЫЗЫВАЕМЫМИ ЗАЯВКАМИ	181
23. Nikol’skii D.N., Krasnov A.E. NETWORK TRAFFIC PREPARATION FOR ITS STATES ANALYSIS BY THE AGGREGATED DATA PACKETS PARTIAL CORRELATIONS METHOD	188
24. Nekrasova R. S., Morozov E. V., Efrosinin D. V. STABILITY ANALYSIS OF AN UNRELIABLE TWO-CLASS RETRIAL SYSTEM WITH CONSTANT RETRIAL RATES.....	194
25. Зверкина Г.А., Кошелев А.А. О МЕТОДЕ МОДЕЛИРОВАНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ С ПОМОЩЬЮ ЕЁ ИНТЕНСИВНОСТИ	200
26. Daraseliya A.V., Sopin E.S. ON THE ANALYSIS OF A RESOURCE LOSS SYSTEM WITH THE WAITING BUFFER	206
27. Orlov Yu. N., Voronina M. Yu THE ERROR CORRECTION METHOD IN THE PROBLEM OF AUTOMATIC AUTHORSHIP IDENTIFICATION OF LITERARY TEXT	212
28. Зорин А.В. О ПЕРИОДЕ ЗАНЯТОСТИ И ЗАГРУЗКЕ СИСТЕМЫ ОБСЛУЖИВАНИЯ С РАЗДЕЛЕНИЕМ ВРЕМЕНИ В СЛУЧАЙНОЙ СРЕДЕ	228
29. Рындиг А.В., Туренова И.А., Моисеева С.П., Пакулова Е.А. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ДВУХПОТОКОВОЙ СИСТЕМЫ ПЕРЕДАЧИ ДАННЫХ	234

30. Бушкова Т.В., Моисеева С.П.	ГАУССОВСКАЯ АППРОКСИМАЦИЯ ДЛЯ РЕСУРСНОЙ ГЕТЕРОГЕННОЙ СМО $(GI + 2M)(2,N)/GI(2)/\infty$	240
31. Пауль С.В., Шульгина К.С., Лизюра О.Д., Шашев Д.В.	ИССЛЕДОВАНИЕ ЦИКЛИЧЕСКИХ СИСТЕМ С ПОВТОРНЫМИ ВЫЗОВАМИ В КЛЮЧЕ ПОСТРОЕНИЯ СЕТЕЙ ПЕРЕДАЧИ ДАННЫХ	247
32. Paul S.V., Nazarov A.A., Phung-Duc T., Morozova M.A.	ANALYSIS OF TANDEM RETRIAL QUEUE WITH COMMON ORBIT AND MMPP INCOMING FLOW	255
33. Кудрявцев Е.В., Федоткин М.А.	ИЗУЧЕНИЕ ПРОЦЕССА АДАПТИВНОГО УПРАВЛЕНИЯ КОНФЛИКТНЫМИ ПОТОКАМИ КОКСА-ЛЬЮИСА ПУТЕМ ИМИТАЦИОННОГО МОДЕЛИРОВАНИЯ	262
34. Zverkina G.A.	ON ASYMPTOTIC ANALYSIS OF QUASI-REGENERATIVE PROCESSES	268
35. Пешкова И.В.	О ПРЕДЕЛЬНОМ РАСПРЕДЕЛЕНИИ МАКСИМУМА СТАЦИОНАРНОГО ВРЕМЕНИ ОЖИДАНИЯ В $GI/G/1$ С ЭКСПОНЕНЦИАЛЬНЫМ-ПАРЕТО ОБСЛУЖИВАНИЕМ	274
36. Moshnikov A.	COMPARISON OF APPROACHES TO COMPONENT RELIABILITY ALLOCATION FOR DISTRIBUTED CONTROL SYSTEMS	279
37. Vasilyev S.A., Tsareva G.O., Bouatta M.A.	NUMERICAL ANALYSIS OF LARGE-SCALE QUEUEING SYSTEM WITH A SMALL PARAMETER	285
38. Lukashenko O.V.	ON THE RELIABILITY ESTIMATION OF THE FBM MULTI-PHASE DEGRADATION SYSTEM	291
39. Bulinskaya E.V.	STABILITY OF SOME APPLIED PROBABILITY MODELS	297
40. Keyela P., Yartseva I.S., Gaidamaka Yu.V.	ANALYTICAL MODEL OF DATA TRANSMISSION THROUGH NARROWBAND-IOT TECHNOLOGY	304
41. Астафьев С.Н.	РАСЧЕТ МАТРИЧНО-АНАЛИТИЧЕСКОЙ МОДЕЛИ СУПЕРКОМПЬЮТЕРА В ПЕРЕХОДНОМ РЕЖИМЕ	310
42. Orlov D.A.	ON APPLICATION OF SOURCE CODE ANALYSIS TECHNIQUES TO HTML PAGES DATA EXTRACTION	316
43. Adou K.Y.B., Markova E.V., Gaidamaka Yu.V.	QUEUEING SYSTEM FOR ANALYZING THE OPERATION OF 5G NETWORK WITH NS UNDER PREEMPTION-BASED SCHEDULER	322
44. Razumchik R.V., Meykhanadzhyan L.A., Pyatkina D.A.	EXISTENCE OF STATIONARY QUEUE-SIZE DISTRIBUTIONS IN THE SYSTEMS THAT WORK ONLY ON THE BIGGEST BATCHES OF CUSTOMERS	328

45. Bobrikova E. V., Platonova A. A., Medvedeva E. G., Gaidamaka Yu. V., Shorgin S. Ya.	A MACHINE LEARNING APPROACH FOR PREDICTING SINR	333
46. Sagdatullin A.	STATE OBSERVER SYSTEM BASED ON K-MEANS CLUSTERING MACHINE LEARNING MODEL FOR CYBER-SECURITY OF INDUSTRIAL NETWORK	339
47. Ivanova N.M.	RELIABILITY ANALYSIS OF A K-OUT-OF-N SYSTEM IN CASE OF FULL REPAIR AFTER ITS FAILURE	345
48. Shukmanova A.A., Yermakov A.S., Paltashev T.T., Mamyrova A.K.	CLOSED STOCHASTIC NETWORK OF THE NEEDHAM-SCHROEDER MODEL FOR OIL PIPELINE DATA TRANSMISSION	352
49. Широков В. Л.	ИНДУСТРИЯ 4.0, СИСТЕМА MESH, МОДЕЛИ И ЭКОСИСТЕМА NG СЕТЕЙ	360
50. Botvinko A.Yu., Samouylov K.E.	FIREWALL SIMULATOR DEVELOPMENT FOR PERFORMANCE EVALUATION OF RANGING A FILTRATION RULES SET	372
51. Абрамян В.Л., Ларинов А.А.	ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ ВЕРОЯТНОСТИ ИДЕНТИФИКАЦИИ RFID-МЕТКИ С ПОМОЩЬЮ RFID-СЧИТЫВАТЕЛЯ, РАЗМЕЩЕННОГО НА БПЛА	386
52. Zirak Q., Shashev D.V.	COLLISION PROVENANCE USING DECENTRALIZED LEDGER AS A BLOCKCHAIN/HASHGRAPH IN SWARM OF DRONES	394
53. Dudin A.N., Dudin S.A., Dudina O.S.	RETRIAL QUEUING SYSTEM WITH LIMITED PROCESSOR SHARING DISCIPLINE	402
54. Клименок В.И., Дудин А.Н.	О РАСПРЕДЕЛЕНИИ ЧИСЛА ПОДРЯД ПОТЕРЯННЫХ ЗАПРОСОВ В СИСТЕМЕ MAP/P H/1/N	408
55. Kukunin D.S., Berezkin A.A., Kirichek R.V.	CODE DIVISION BASED ON M-SEQUENCES AND ITS OPTIMIZATION	415
56. Berezkin A.A., Kukunin D.S., Slepnev A.V., Kirichek R.V.	EFFICIENT DATA CODING METHODS BASED ON NEURAL NETWORKS	421
57. Brekhov O.M. and Klimenko A.V.	THE ESTIMATION OF MICROCHIP TESTING PROCESS DURATION BASED ON EXTENDED FAULT INJECTION METHOD	427

УДК: 519.718

Требования сценариев URLLC, mMTC, eMBB к уровням L0, L1 транспортной сети IMT-2020/5G

Д.А. Аминев¹, Е.Г. Богданова³, Д.В. Козырев^{1,2}

¹Институт проблем управления им. В.А. Трапезникова РАН, ул. Профсоюзная, 65, Москва, Россия

²Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия

³ООО «Т8 НТЦ» Краснобогатырская ул., д. 44 стр.1, Москва, 107076, Россия

aminev.d.a@ya.ru, bogdanova@t8.ru, kozyrev-dv@rudn.ru

Аннотация

Показана актуальность сценариев для сетей мобильной связи 5G. Раскрыта архитектура физического и канального уровней транспортной сети 5G. Проведён анализ требований сценариев eMBB, URLLC, mMTC к физическому и канальному уровням транспортной сети, их ориентировочная численная оценка. Выявлены требования, оказывающие доминантное влияние при построении сети, введена их формализация.

Ключевые слова: транспортная сеть, формализация требований, тракт передачи данных, мобильная связь, 5G.

1. Введение

Переход к стандарту 5G приводит к изменениям всех составляющих сетей мобильной связи: сети радиодоступа (RAN), ядра (Core), единой плоскости управления и транспортной сети, обеспечивающей соединение всех компонентов в единое комплексное решение. Три категории сценариев применения, определенные стандартом 3GPP для сетей 5G/IMT-2020, оказывают существенное влияние на требования, предъявляемые ко всем этим компонентам и транспортной сети в частности:

- сверхширокополосная мобильная связь eMBB (enhanced Mobile Broadband);
- сверхнадежная межмашинная связь с низкими задержками URLLC (Ultra-Reliable Low Latency Communication);
- массовая межмашинная связь mMTC (Massive Machine-Type Communications).

Сценарий eMBB предполагает большую пропускную способность сети, необходимую для поддержки высоких скоростей передачи данных мобильных абонентов. Сценарий URLLC накладывает существенные ограничения на задержки и предъявляет повышенные требования к пропускной способности и надежности сетевого оборудования. Сценарий mMTC предполагает соединение большого количества движущихся объектов и предъявляет требования к гибкости сети и доступности соединений.

Каждый конкретный сценарий из трёх категорий предусматривает ряд требований, выполнение которых будет определять оптимальную архитектуру сети. При этом архитектура должна быть масштабируема, чтобы поддерживать пользовательские сценарии не только на начальном этапе внедрения технологии, но и в дальнейшем, при развитии сценариев eMBB, URLLC, mMTC. Проведение такой оптимизации математическими методами невозможно без предварительной формализации требований сценариев.

2. Архитектура транспортной сети IMT-2020/5G

На рисунке 1 представлена архитектура транспортной сети с декомпозицией функциональных элементов базовой станции в соответствии с рекомендациями 3GPP [1, 3]. Между основными компонентами RU (Remote Unit), DU (Distributed Unit), CU (Central Unit) выделяются следующие участки: Fronthaul, Midhaul и Backhaul. Сегменту Fronthaul соответствует домен доступа в терминологии классической транспортной сети; сегменту Midhaul — домен метро-агрегации; сегменту Backhaul — домен магистральной транспортной сети [2-4]. Наличие или отсутствие сегментов Fronthaul и Midhaul определяется сценарием развертывания сети радиодоступа.

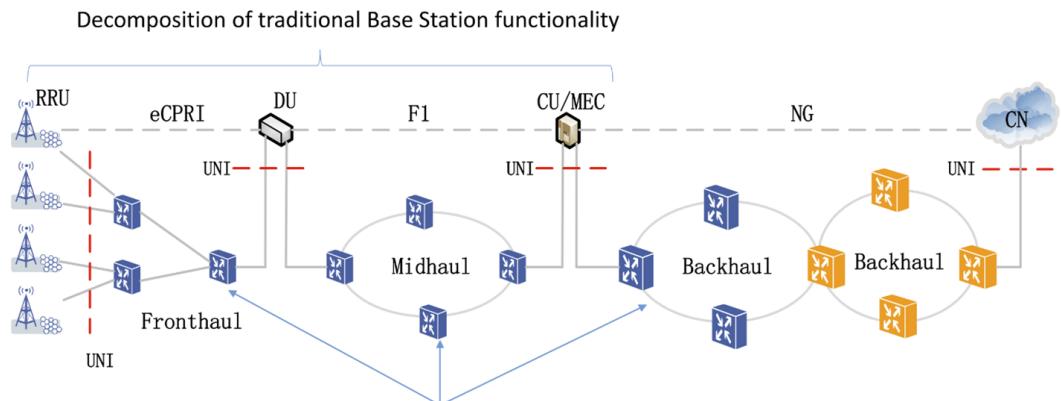


Рис. 1. Архитектура транспортной сети для IMT-2020/5G [1]

Сеть IMT-2020/5G рассматривается стандартами как пакетная сеть, где пакетные коммутаторы и маршрутизаторы IP/Ethernet организуют связность основных компонентов сети 5G на уровне L2/L3 модели межсетевого взаимодействия OSI [5]. Эти соединения организуются поверх физической инфраструктуры L0, например, в темном волокне или с использованием одной из технологий физического уровня — PON, xWDM, беспроводная связь.

На уровне L1 между пакетной и физической сетями может использоваться транспортная технология, например, OTN (Optical Transport Network), которая обеспечивает агрегацию и прозрачную передачу больших объемов трафика, а также предоставляет ряд преимуществ при организации сквозных соединений 5G. На рисунке 2 представлена логическая модель транспортной сети, объединяющая в себе технологии разных уровней.

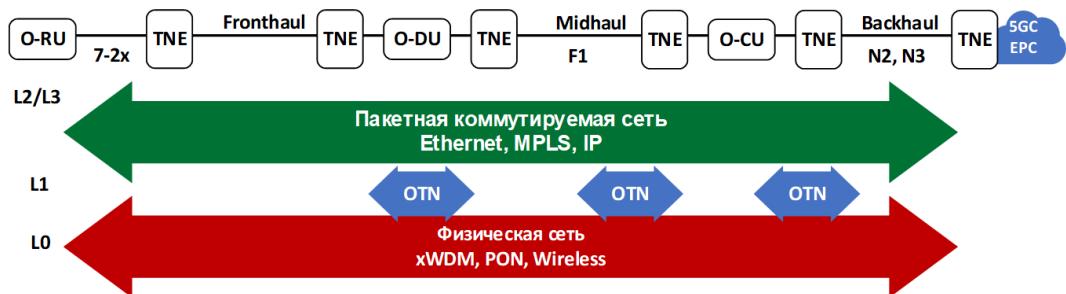


Рис. 2. Логическая модель транспортной сети (общий случай)

В рамках статьи будут рассмотрены требования к двум нижним уровням L0, L1, а именно к оборудованию физического уровня xWDM и транспортному оборудованию OTN. Устройства коммутации/маршрутизации L2/L3, реализующие передачу пакетного трафика IP/Ethernet, является клиентским оборудованием и рассматривается в статье только с точки зрения требований к интерфейсам взаимодействия.

3. Анализ и формализация требований

Опираясь на указанные категории пользовательских сценариев можно выделить общие требования к транспортной сети: высокая пропускная способность с возможностью масштабирования, низкие задержки и передача сигналов синхронизации, надёжность, а также гибкость сети и доступность соединений в любой момент времени. Выполнение каждого из указанных требований может быть обеспечено реализацией различных программных и аппаратных функций.

3.1. Масштабирование пропускной способности. Согласно аналитическим исследованиям [5] в сетях 5G прогнозируется десятикратное увеличение пропускной способности и количества абонентских устройств по сравнению с предыдущим поколением мобильной связи. В связи с этим транспортное оборудование на каждом сегменте должно обеспечить высокую пропускную способность сети IMT-2020/5G и возможность её масштабирования по мере увеличения абонентской нагрузки.

В общем случае пропускная способность Fronthaul для одного сайта зависит от количества секторов антенны, пропускной способности радиоканала на каждой несущей, реализации функционала MIMO для каждой несущей. В спецификации O-RAN [7] приводится оценка трафика Fronthaul для разных сценариев, для опции 7.2x функционального разделения базовой станции:

- малые соты: один сектор, несущие в диапазоне либо mmWave, либо Sub6 с низким порядком MIMO;
- средние соты: много секторов, несущие как в Sub6-, так и в mmWave-диапазонах со средним порядком MIMO;
- большие соты: много секторов, несущие в обоих диапазонах, mmWave- и Sub 6-, с использованием Massive MIMO.

Опция 7.2x предполагает, что оборудование радиодоступа на сегменте Fronthaul после обработки трафика между базовыми станциями DU/CU и блоками обработки RU формирует потоки eCPRI, которые передаются через стандартные интерфейсы L2/L3 10GE/25GE. Это означает, что оборудование оптического транспорта должно поддерживать клиентские сигналы 10GE/25GE.

Требуемая ёмкость на сегменте Fronthaul может быть достигнута следующими способами:

1) *Организация тёмных волокон для соединений RU и DU по принципу точка-точка (L0)* — в каждом волокне возможна передача трафика одной антенны со скоростью 10 Гбит/с или 25 Гбит/с в зависимости от интерфейса. Транспортный уровень L1 отсутствует.

2) *Мультиплексорование каналов на физическом уровне (L0)* — повышается эффективность использования волокна за счёт использования оптических xWDM-мультиплексоров. В одной паре волокон можно передать более 80 каналов по 10 или 25 Гбит/с. Как и в первом случае, транспортная технология уровня L1 отсутствует.

3) *OTN-агрегация* — агрегировать большое количество клиентских потоков можно с помощью технологии оптического транспорта OTN (на уровне L1). Для еще более эффективного использования волокна OTN-агрегаторы (мукспондеры) могут подключаться к портам xWDM-мультиплексоров.

Формализуем требования к пропускной способности и определим численные критерии масштабирования ёмкости транспортной инфраструктуры для трёх сценариев. Определим следующие переменные:

N_{fiber} — число доступных пар волокон на сайте оператора,

N_{10G}/N_{25G} — требуемое число соединений со скоростью 10 Гбит/с / 25 Гбит/с,

N_{mux} — число доступных портов оптических мультиплексоров,

$N_{OTN_10G}, N_{OTN_25G}^*$ — число доступных клиентских интерфейсов 10GE и 25GE OTN-агрегаторов.

Для сценариев 1-3 масштабируемость пропускной способности Fronthaul-сегмента B_{scal} зависит от доступности транспортных ресурсов и от потенциальной возможности увеличения ёмкости до максимального значения на конкретном Fronthaul-сегменте.

Доступность ресурсов определим коэффициентами k_{fiber} , k_{mux} или k_{OTN} для сценариев 1-3 соответственно. Возможность увеличения ёмкости до максимального значения B_{max} для конкретного Fronthaul-сегмента определяется коэффициентом b_{free} .

$$b_{free} = \frac{(B_{max} - N_{10G} * 10 - N_{25G} * 25)}{B_{max}}.$$

Масштабируемость емкости B_{scal} для разных архитектур можно оценить следующим образом:

1) При использовании P2P-соединений:

$$B_{scal} = k_{fiber} * b_{free}, k_{fiber} = \frac{N_{fiber}}{N_{10G} + N_{25G}},$$

где k_{fiber} — коэффициент доступности волокна:

2) При использовании оптических мультиплексоров

$$B_{scal} = k_{mux} * b_{free}, k_{mux} = \frac{N_{mux} * N_{fiber}}{N_{10G} + N_{25G}},$$

где k_{mux} — коэффициент доступности портов оптического мультиплексора:

3) При использовании OTN-агрегации

$$B_{scal} = k_{OTN} * b_{free}, k_{OTN} = \frac{(N_{OTN_{10G}} + N_{OTN_{25G}}) * N_{mux} * N_{fiber}}{N_{10G} + N_{25G}},$$

*Число доступных клиентских интерфейсов N_{OTN_10G}, N_{OTN_25G} приводится для сценария максимально эффективного использования портов OTN-мукспондера, когда максимальное количество портов используется для 25GE-клиентов, а оставшаяся ёмкость устройства занимается клиентами 10GE. При этом, выполняется условия $N_{OTN_10G} \geq N_{10G}$, $N_{OTN_25G} \geq N_{25G}$.

где k_{OTN} – коэффициент доступности клиентских интерфейсов OTN-мукспондера.

Абсолютные значения максимальной пропускной способности Backhaul и Midhaul на сегодняшний день не определены в стандартах и спецификациях. Предполагая, что Midhaul и Backhaul реализованы на OTN/DWDM-технологии, требование к пропускной способности транспортного оборудования на данных участках всегда будет выполнено.

3.2. Временные характеристики сети: задержки и поддержка передачи сигналов синхронизации. Ключевыми отличиями транспортной сети 5G от классических транспортных сетей являются жесткие требования к задержкам и к поддержке передачи сигналов синхронизации. Сеть 5G является пакетно-ориентированной и не может без дополнительных механизмов обеспечить выполнение этих требований. Значения задержки и джиттера не являются детерминированными для пакетных коммутаторов и маршрутизаторов. Только применение специальных механизмов TSN (TSN – Time Sensitive Networking) позволяет обеспечить предсказуемые временные характеристики.

Задержка на оптическом уровне является детерминированной. Она обусловлена, в первую очередь, задержкой распространения сигнала в оптоволокне (5 мкс/км). В OTN/DWDM самый большой вклад в задержку вносит процедура исправления ошибок FEC, а также наличие буферов для обработки заголовков OTN в процессорах. Задержки $T_{L0/L1}$ и $T_{L2/L3}$ на разных уровнях в дальнейшем будем рассматривать по-отдельности. Величина сквозной задержки потока данных на сегодняшний день специфицирована для нескольких пользовательских сценариев. Обозначим эту величину как T_{E2E} . Рисунок 3 отражает распределение задержки по сегментам сети для URLLC и eMBB в соответствии с рекомендациями ITU-T [8, 9] и 3GPP [10][†]. Значения задержек для трафиков плоскостей данных (U, User Plane) и управления (C, Control Plane) не всегда совпадают, так как эти потоки могут терминироваться в разных элементах сети 5G.

Наиболее критичным к задержкам сегментом является Fronthaul. Расстояние между RU и BBU не должно превышать 10-20 км, что соответствует задержке 100-200 мкс при двунаправленной передаче и определяется ограничениями протоколов CPRI/eCPRI, которые должны поддерживаться транспортным оборудованием. Точные значения допустимой задержки между точками терминации интерфейсов CPRI и eCPRI приведены в таблице 1 и обозначены как T_{E2E_FH} .

Кроме допустимого абсолютного значения задержек на сегменте Fronthaul необходимо учитывать асимметрию задержек по направлениям передачи (вверх – от RU до DU, вниз – от DU до RU). Ограничения связаны с синхронизацией фазы и времени между элементами базовой станции. Протоколы CPRI и eCPRI

[†]На сегодняшний день требования к задержкам для сценария mMTC стандартами не определены.

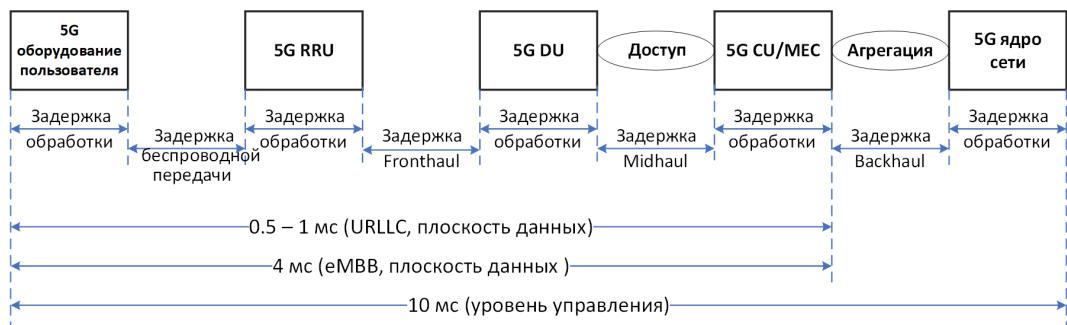


Рис. 3. Распределение требуемых задержек на сети 5G для сценариев URLLC и eMBB для плоскости данных и плоскости управления

Таблица 1. Требования протоколов CPRI и eCPRI к транспортной сети

Требование к транспортному оборудованию FH	Допустимые значения при использовании CPRI	Допустимые значения при использовании eCPRI
Допустимая двусторонняя задержка T_{E2E_FH}	От 50 до 250 мкс, в зависимости от мобильного приложения	100 мкс – 100 мс в зависимости от класса трафика
Допустимая ошибка времени (TE)	± 16 нс	± 10 нс... $\pm 1,36$ мкс в зависимости от класса трафика
Точность установки частоты (FE)	50 ppb	16 ppb
Точность установки фазы (PE)	1,5 нс	1,5 нс

предъявляют разные требования к точности. В общем случае асимметрия задержки на Fronthaul не должна превышать значения ошибки времени TE (Time Error) (Таблица 1). Асимметрия задержки в DWDM-оборудовании определяется следующими факторами:

- разная оптическая длина волокон при двунаправленной передаче;
- наличие компенсаторов дисперсии в DWDM-сети с некогерентными транспондерами;
- OTN/DWDM-транспондеры, содержащие в составе FIFO-буферы.

Дополнительно ошибка времени на сети возникает в IP-маршрутизаторах и Ethernet-коммутаторах.

В транспортном оборудовании возможна компенсация асимметрии задержки как на электрическом уровне (L1) при обработке сигналов, так и на физическом

уровне (L0). Возможно исключение асимметрии задержки с помощью архитектурных решений, например, организацией двунаправленной передачи по одному оптическому волокну (т.н. BiDi-решения).

В системах 5G выдвигаются строгие требования к синхронизации фазы/времени между блоками радиооборудования. Сигналы синхронизации доставляются от первичного источника синхронизации до радиоблоков RU по транспортной сети. Протоколы сетевой синхронизации работают на уровне L2 – это т.н. синхронный Ethernet (SyncE) и протокол точного времени IEEE 1588-2008 (PTP). Оптическая сеть xWDM предоставляет физический уровень для прозрачной передачи Ethernet-трафика, в составе которого передаются пакеты 1588 и SyncE, и должна обеспечивать требуемую точность установки частоты, времени и фазы.

Формализуем требования к времененным характеристикам и определим численные критерии, необходимые для включения данного целевого параметра в модель оптимальной транспортной сети. Введём переменные:

$T_{L0/L1}$ – суммарное значение задержек (TFIBER, TFEC, TOTN) во всех элементах оборудования L0/L1 при прохождении сигнала от RU до ядра сети 5G CN (рисунок 3, является фиксированным для данной конфигурации сети).

$T_{L2/L3}$ – суммарное значение задержек во всех элементах оборудования L2/L3 при прохождении сигнала от RU до ядра сети 5G CN (рисунок 3, не является константой для данной конфигурации сети и зависит от загрузки).

T_{E2E_FH} – суммарное значение задержек во всех элементах оборудования L0/L1 на сегменте Fronthaul;

T_{E2E_max} – максимально допустимая задержка на всю сеть (рисунок 3), включает в себя все источники задержки при передаче сигнала «из конца в конец».

$T_{E2E_FH_max}$ – максимально допустимая задержка на сегменте Fronthaul.

TE, \bar{FE}, \bar{PE} – ошибки времени, установки частоты и фазы соответственно на сегменте Fronthaul.

TE_{max} – максимально допустимая ошибка времени на сегменте Fronthaul;

FE_{max} – максимально допустимая ошибка установки частоты на сегменте Fronthaul;

PE_{max} – максимально допустимая ошибка установки фазы на сегменте Fronthaul.

Целевой параметр k_{time} , характеризующий временные характеристики сети, состоит из слагаемых:

$$k_{E2E} = \frac{T_{E2E_max} - T_{L2/L3} - T_{L0/L1}}{T_{E2E_max}}; k_{FH} = \frac{T_{E2E_FH_max} - T_{E2E_FH}}{T_{E2E_FH_max}};$$
$$k_{TE} = \frac{TE_{max} - TE}{TE_{max}}; k_{FE} = \frac{FE_{max} - FE}{FE_{max}}; k_{PE} = \frac{PE_{max} - PE}{PE_{max}}.$$

Тогда $k_{time} = k_{E2E} + k_{FH} + k_{TE} + k_{FE} + k_{PE}$. Полагаем, что при формировании целевого параметра «задержки» в общем случае каждое из слагаемых имеет оди-

наковый вес, так как все они одинаково важны для соответствия транспортного уровня требованиям к времененным характеристикам.

3.3. Гибкость сети и доступность соединений. Под гибкостью транспортной сети понимается связность узлов на транспортном уровне. Гибкость обеспечивает возможность установления соединения между любыми двумя абонентами и определяется поддержкой коммутации трафика в IP-маршрутизаторах (L3), Ethernet-коммутаторах (L2), OTN кросс-коммутаторах (L1), оптических устройствах коммутации длин волн (L0).

Соединения на сегменте Fronthaul в основном реализованы по принципу точка-точка, и необходимость кросс-коммутации отсутствует. Требование к гибкости и, соответственно, поддержке кросс-коннекта предъявляется на сегментах с более сложной топологией, Midhaul и Backhaul.

Сеть, реализованная на стеке технологий IP/MPLS, обеспечивает коммутацию с высокой гранулярностью на уровне отдельных пакетов. Пакеты данных направляются в нужный порт, соответствующий требуемому направлению, за счет обработки IP-заголовков или специальных MPLS-меток. Однако, эти операции требуют высокой производительности оборудования и, как следствие, высокого энергопотребления. Кроме того, задержки в L2/L3-устройствах зависят от загрузки и не являются детерминированными.

В сложной топологии Midhaul/Backhaul для некоторых устройств L2/L3 большая часть трафика является транзитной, поэтому для увеличения эффективности использования сетевых ресурсов целесообразно организовать т.н. bypass или обход промежуточных коммутаторов и маршрутизаторов. Тогда трафик, минуя L2/L3-обработку, будет проходить по маршруту, сформированному из OTN- или фотонных коммутаторов (рисунок 4а).

Таким образом, преимуществами коммутации L1/L0 являются снижение энергопотребления роутеров, разгрузка ресурсов L2/L3, а также снижение задержки прохождения трафика по маршруту и её детерминированность.

Для организации прозрачных сквозных соединений и управления трафиком на уровне отдельных клиентских сервисов используется оборудование OTN кросс-коммутации с полнодоступной матрицей коммутации.

Если необходимо управление трафиком (перенаправление потоков в промежуточных узлах) на уровне отдельных длин волн, транспортное оборудование должно включать фотонные коммутаторы – оптические мультиплексоры ввода-вывода ROADM (Reconfigurable Optical Add-Drop Multiplexer). ROADM не обеспечивают такой гранулярности, как OTN-коммутаторы, однако позволяют более эффективно управлять большими объемами трафика без оптоэлектронного преобразования (рисунок 4б). Гибкость в данном случае будет определяться количеством направлений в ROADM-узле.

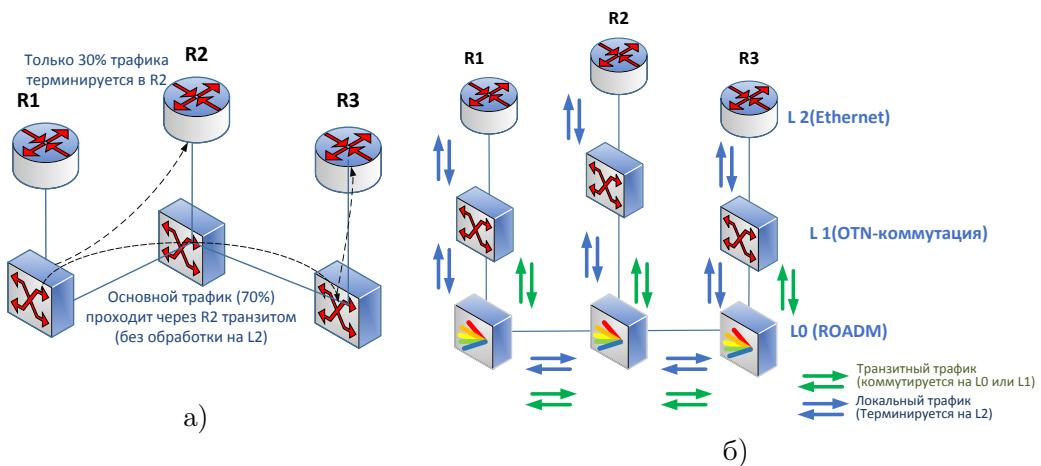


Рис. 4. Обход промежуточных L2/L3-коммутаторов с коммутацией в OTN/DWDM-узлах (а); уровни коммутации в транспортной сети (б) – электронная OTN-коммутация (L1) и коммутация длин волн в ROADM (L0)

Максимальная эффективность оптической (фотонной) коммутации обеспечивается конфигурацией «CDC» (Colorless, Contentionless, Directionless) устройства ROADM, что позволяет переключать трафик из любого входного порта, т.е. с любого направления, в любой выходной порт вне зависимости от пришедшей длины волны. Как и для OTN-коммутаторов назовем это свойство полнодоступностью.

В рамках данной статьи подробно не рассматриваются возможности и характеристики коммутационного оборудования. Отметим только, что возможность обхода L2/L3-узлов по транспортной сети обеспечивается полнодоступной коммутацией на транспортном уровне, т.е. возможностью в любом узле сети любой входной порт, соответствующий определенному направлению, связать с любым выходным портом/направлением.

Для модели оптимальной транспортной сети 5G/IMT-2020 введём характеристику, показывающую возможность обхода коммутаторов/маршрутизаторов на уровне транспортного оборудования, определяемую наличием соединения на транспортном уровне между двумя узлами и ёмкостью, которую обеспечивает это соединение.

Наличие соединения определяется наличием в узле оборудования OTN-кросс-коннекта или ROADM с полнодоступной коммутацией. Ограничение ёмкости на транспортном уровне может быть связано с тем, что число длин волн (на L0) или число OTN-каналов (на L1) меньше, чем число клиентских потоков при пиковой нагрузке. Пиковая нагрузка на сеть имеет место, когда клиенты со всех

направлений единовременно коммутируются на конкретный узел по одному из линейных участков.

Введем коэффициенты k_{L0} и k_{L1} для каждой пары смежных узлов сети. Они показывают отношение доступных соединений на транспортном уровне к максимально возможному количеству клиентских потоков со всех направлений между данными узлами:

$$k_{L0} = \frac{N_{L0}}{N_{L2/L3}}, k_{L1} = \frac{N_{L1}}{N_{L2/L3}}$$

N_{L0} — количество соединений на фотонном уровне (длин волн) для заданной пары смежных узлов, N_{L1} — количество OTN-каналов для заданной пары смежных узлов, $N_{L2/L3}$ — максимальное количество клиентских потоков со всех направлений для заданной пары смежных узлов.

Для упрощения примем, что пропускная способность каждого транспортного канала не меньше требуемой клиенту полосы, поэтому в формуле фигурирует только количество каналов, поскольку OTN обеспечивает масштабируемость ёмкости, а современное оборудование агрегирует и передает клиентские потоки скоростью до 400 Гбит/с.

В модели необходимо учесть k_{L0} и k_{L1} для каждой пары смежных узлов и оценить гибкость сети в целом, поэтому просуммируем полученные коэффициенты и нормируем их к количеству соединений/пар смежных узлов в сети Midhaul или Backhaul M:

$$K_{L0} = \frac{\sum\limits_1^M k_{L0}}{M}, K_{L1} = \frac{\sum\limits_1^M k_{L1}}{M}.$$

Если порты IP-маршрутизаторов ассоциируются с оптической длиной волны и соединения между источником и приёмником устанавливается через ROADM-узлы, сеть характеризуется только коэффициентом K_{L0} . Если используется более гранулярная коммутация, и порты L2/L3 связаны с клиентскими портами OTN-коммутаторов, необходимо использовать коэффициент связности K_{L1} . Физическая связность узлов должна поддерживаться плоскостью управления. Маршруты трафика могут строиться оператором в системе управления транспортным оборудованием или автоматически прокладываться с помощью протоколов GMPLS (Generalized Multiprotocol Label Switching). Вычисление маршрутов в простом случае происходит распределенно в узлах сети или централизованно в SDN-контроллере, если сеть является программно-управляемой.

3.4. Надёжность соединений. Надёжность сети определяется доступностью соединений в любой момент времени и механизмом реализации защитных переключений.

В сети IMT-2020/5G защитные переключения могут быть организованы на разных уровнях: в оборудовании L2/L3, на электрическом уровне L1, а также на фотонном уровне L0. Возможность переключения на резервный маршрут тесно связана с функционалом коммутации, описанным в предыдущем разделе, а также с возможностями плоскости управления.

К стандартизованным механизмам защитного переключения относятся схемы: 1+1 OLP (Optical Line Protection), 1+1 OMSP (Optical Multiplex Section Protection) или E-SNCP (Electrical SubNetwork Connection Protection) [11,12]. OLP работает на уровне L0, в то время, как OMSP и E-SNCP реализуются в OTN-оборудовании на L1.

Расширить возможности аппаратного резервирования позволяет плоскость управления, если она реализована с использованием группы протоколов ASON/GMPLS (Automatically Switched Optical Network/ Generalized Multiprotocol Label Switching). Становится возможным автоматическое восстановление трафика. Специальные протоколы автоматически определяют топологию, обнаруживают доступные сетевые ресурсы и резервируют их на случай аварийного переключения. В случае отказа, за 50 мс происходит переключение на альтернативный маршрут. При этом маршруты формируются на основе метрик, определяемых оператором или приложением.

Доступность соединений при IP-коммутации оценивается значением $\sim 99.9\%$, в то время, как доступность в системе передачи на фотонных коммутаторах $\sim 99.999\%$. Общим требованием к доступности в сети 5G является значение 99.999%. Применение ASON/GMPLS увеличивает доступность соединений до значения 99.9999%, т.к. позволяет защитить трафик даже при двойных отказах.

В таблице 2 приведено сравнение защитных механизмов.

Внутри каждого класса в таблице 2 можно выделить разные типы защиты. Однако на данном этапе в модели будем учитывать только факт наличия такого механизма на одном из уровней, L0 или L1. Введём целевой параметр P . Тогда, если на уровне Control Plane реализован стек GMPLS-протоколов, обеспечивающий более эффективные механизмы защиты, целевой параметр P принимает максимальное значение, равное 1.

Если механизм защитного переключения управляет только ресурсами системы управления: 1+1 OLP, 1+1 OMSP или E-SNCP [11, 12], целевой параметр защиты имеет меньшее значение $P = 0.5$. Возможна ситуация, когда защитное соединение не предусмотрено (например, транспондеры в сети имеют только один линейный выход и оптических блоков резервирования не предусмотрено). В таком случае $P = 0$.

Таблица 2. Сравнение защитных механизмов L2/L3, L1, L0

Характеристика защитного переключения	Ethernet, IP/MPLS (L2/L3)	OTN (L1)	Оптический уровень (L0)
Автоматизация	Автоматические переключения	В системе управления или автоматически при реализации GMPLS	В системе управления или автоматически при реализации GMPLS
Уровень защиты	Защищаются пакеты данных, Защита только трафика с высоким приоритетом	Защита отдельных клиентов или оптического канала (длины волны)	Защита всего линейного тракта или отдельного волокна
Необходимость терминации (обработки трафика) в узле	Терминация в L2/L3-устройствах — обработка пакетов	Обработка заголовков OTN в узлах OTN-XC	Потоки не терминируются, разбор трафика не требуется
Время восстановления соединения	Для IP – минуты, Для IP/MPLS <1с	<50 мс	<50 мс
Доступность соединений	99,9%	Без GMPLS – 99,999%; С GMPLS – 99,9999%	Без GMPLS – 99,999% С GMPLS – 99,9999%

4. Заключение

На основе анализа общих требований к транспортной сети IMT-2020/5G выявлены основные требования к физическому и канальному уровням: высокая пропускная способность с возможностью масштабирования, низкие задержки и передача сигналов синхронизации, надёжность, а также гибкость сети и доступность соединений в любой момент времени.

В результате формализации введены следующие параметры:

- B_{scal} характеризует возможность масштабирования пропускной способности сети;
- k_{time} характеризует временные характеристики;
- K_{L0}/K_{L1} характеризует гибкость транспортной сети на уровнях L0 или L1;
- P — наличие механизмов защиты соединений.

Эти требования станут целевыми параметрами в решении задачи построения оптимальной архитектуры на этапе проектирования транспортной сети, а также при построении оптимальных маршрутов передачи трафика на реализованной

сети в зависимости от сценария (eMBB, URLLC, mMTC). При оптимизации архитектуры в зависимости от выбранного сценария формализованные требования будут учтены с соответствующими весовыми коэффициентами w . Так, например, для сценария URLLC критичным параметром является задержка TL0/L1, поэтому параметр k_{time} будет иметь наибольший вес. Для сценария eMBB важна пропускная способность, а значит, максимальный вес в модели будет иметь параметр B_{scal} . Сценарий mMTC предполагает множественные динамические соединения перемещающихся объектов, и вес K_{L0}/K_{L1} будет наиболее значим.

Литература

1. 3GPP Technical specification. NG-RAN; Architecture description.
2. Богданова Е. Транспортная сеть 5G/IMT-2020// Первая миля. №7. С.40-47. - 2019.
3. ITU-T Recommendation G.8300: Characteristics of transport networks to support IMT-2020/5G.
4. ITU-T Technical Report. Transport network support of IMT-2020/5G.
5. 3GPP Technical specification. NG-RAN; NG data transport.
6. Light Reading. “5G Network & Service Strategies. Operator Survey”, 2021
7. O-RAN.WG9.WTRP-Req-v.01.00. Technical specification. O-RAN Open X-haul Transport Working Group 9. Xhaul Transport Requirements.
8. ITU-T Recommendation G.8271.1: Network limits for time synchronization in packet networks with full timing support from the network»
9. ITU-T Recommendation G.8273.2: Timing characteristics of telecom boundary clocks and telecom time slave clocks for use with full timing support from the network.
10. 3GPP Technical report. Study on scenarios and requirements for next generation access technologies.
11. Recommendation ITU-T G.709/Y.1331. Interfaces for optical transport network
12. Recommendation ITU-T G.831.1 (10/2017). Optical transport network: Linear Protection.

UDC: 519.24

Estimation of the Tail Index of PageRanks in Random Graphs*

Natalia M. Markovich¹ and Maksim S. Ryzhov¹

¹V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences ,

Profsoyuznaya Str. 65, 117997 Moscow , Russia

markovic@ipu.rssi.ru, nat.markovich@gmail.com,maksim.ryzhov@frtk.ru

Abstract

Superstar nodes to which a large proportion of nodes attach in the evolving graphs are considered. We attract results of the extreme value theory regarding sums and maxima of non-stationary random length sequences to predict the tail index of the PageRanks and Max-linear models as influence measures of superstar nodes. To this end, the graphs are divided into mutually weakly dependent communities. Maxima and sums of the PageRanks over communities are used as weakly independent block-data. Tail indices of the block-maxima and block-sums and hence, of the PageRanks and the Max-linear models are found to be close to the minimum tail index of series of representative nodes taken from the communities. The graph evolution is provided by a linear preferential attachment. The tail indices are estimated by data of simulated and real temporal graphs.

Keywords: Random graph, tail index, PageRank, Max-linear model, superstar node, community, evolution, preferential attachment

1. Introduction

Random graphs attract the attention of researchers due to numerous applications including complex networks and communication systems. The randomness of such graphs consists in the random numbers of in-coming and outgoing links of the nodes that are called in- and out-degrees, respectively. Random graphs are subject to heterogeneity of distributions of node indices and their dependence structure. A well-known feature of random graphs such as Web graphs is that the in- and out-degrees are power law distributed. Despite the in- and out-degrees are discrete

*The reported study was funded by the Russian Science Foundation RSF, project number 22-21-00177 (recipient N.M. Markovich, conceptualization, methodology development, formal analysis, writing—original draft preparation; recipient M.S.Ryzhov, software, data validation).

random variables (r.v.s), their distribution can be approximated by regularly varying heavy-tailed distributions [1]. The distribution tail of a non-negative r.v. X is called regularly varying with the tail index (TI) α if it holds $\bar{F}(x) = P\{X > x\} = x^{-\alpha}\ell(x)$, where $\ell(x)$ is a slowly varying function, i.e. $\lim_{x \rightarrow \infty} \ell(tx)/\ell(x) = 1$ holds for any $t > 0$. The smaller TI the heavier the distribution tail. Dependence between nodes The PageRank (PR) is a more general measure of the node influence than the in-degree. The PRs of a random Web page are derived to be regularly varying distributed [1, 2]. Empirical studies stated that the TIs of the in-degrees and PRs of Web pages have a value α smaller than 2. By properties of regularly varying distributions, this means that the variance of the node indices is infinite. Another measure of the node influence is the Max-Linear Model (MLM) [3].

Our aim is to evaluate TIs of the PRs and MLMs of *superstar nodes* connected by a large number of in-coming links to other nodes in evolving random graphs. The edge and node structure of such graphs is changed in time. We use results of the extreme value theory regarding sums and maxima of non-stationary random length sequences to find the latter TIs that is a novelty. To this end, we divide the evolving random graphs into communities by a Directed Louvain's algorithm [4] which constitute weakly dependent subgraphs and take representative nodes from each community. A community consists of nodes that are strongly connected with each other and weakly connected with nodes from other communities [5]. The node indices in the communities may also be dependent and non-stationary distributed. We estimate the TIs of the block-maxima and block-sums of the node PRs over the communities and compare it with the minimum TI of the PRs among the representative series. The graph evolution is provided by a linear preferential attachment (PA) [6]. We study simulated evolving graphs with stationary distributed in- and out-degrees and homogeneous dependence structure and real temporal graphs. The Hill's and QQ plot estimators (see, [7]) to estimate the TI are used.

In Sect. 2.1 we recall theoretical results obtained in [8, 9] and discuss their application to find the TIs of the PRs and MLMs of the superstars. In Sect. 3.1, 3.2 TIs are estimated by data of simulated stationary graphs evolved by the PA and of real temporal graphs. We finalize with conclusions in Sect. 4.

2. Theory and its interpretation for random graphs

2.1. Tail index of random sums and maxima . In [1, 2], the PR R of a randomly chosen Web page (i.e. a vertex of a Web graph) is presented as a solution of a fixed-point problem

$$R =^D \sum_{j=1}^N A_j R_j + Q, \quad (1)$$

assuming that $\{R_j\}$ are independent identically distributed (i.i.d.) copies of R and $E(Q) < 1$ holds. $=^D$ denotes the equality in distribution. The PRs of nearest neighbors of an underlying node which have in-coming links to this node form the random sum in (1) by the original definition of the PR given in [10]. It is stated in [1, 2] that the stationary distribution of R is regularly varying and its TI is determined by the most heavy-tailed distributed term in the regularly varying distributed pair (N, Q) . By replacing the sum by maximum one can obtain similar results with regard to the MLM that is the solution of the following equation

$$R =^D \left(\bigvee_{j=1}^N A_j R_j \right) \vee Q. \quad (2)$$

In [8, 9] the TI of sums and maxima of random length weighted non-stationary sequences was found. The latter sequences can be considered as the rows of a doubly-indexed array of r.v.s ($Y_{n,i} : n, i \geq 1$) in which the "row index" n corresponds to time, and the "column index" i corresponds to the level. The "column" series are assumed to be stationary distributed with regularly varying tails and the TIs $\{k_i\}_{i \geq 1}$. One of the "column" series is assumed to have a minimum TI k_1 . An arbitrary dependence between "column" series is allowed. It was found that the TI of both sums and maxima over rows is equal to k_1 [8]. The same may be true if there are a random number of such the most heavy-tailed "column" series with k_1 [9]. The results remain true if the TIs of elements in the "columns" are different, apart from those "columns" with k_1 [9].

2.2. Random sums and maxima in random graphs . The results in [8, 9] can be applied to the sums and maxima in the right-hand sides of (1) and (2). This allows to find TIs of the PRs and the MLMs of superstar nodes within communities, to which a large proportion of nodes of the community attach. In our empirical study we exclude isolated nodes from the consideration. The PR or the MLM of a superstar node is calculated by sum or maximum of PRs of its nearest neighbors having in-coming links to the latter node. The superstar may have a few followers in other communities but we disregard them. By theory and our simulation study, the TI of their PRs (or the MLMs) may be approximated by the TI of the most heavy-tailed representatives of communities, i.e. by the minimum TI of the representative series. In terms of graphs the random length "row" sequences can be considered as communities of nodes. The "column" series consist of nodes that are taken from each community as their representatives, see Fig. 1.

As the PRs in the communities may be non-stationary distributed and arbitrary dependent that is natural for practice, the representative series with the minimum

TI must be stationary distributed. The latter property may not be fulfilled in heterogeneous random graphs. One has to test the stationarity of the representative series with the minimum TI. The representative series may be formed by branches of all possible trees rooted at each node of the communities. To reduce a number of the representative series we take the i th maxima of each community as a member of the i th representative series. The first maxima are excluded, i.e. $i \geq 2$.

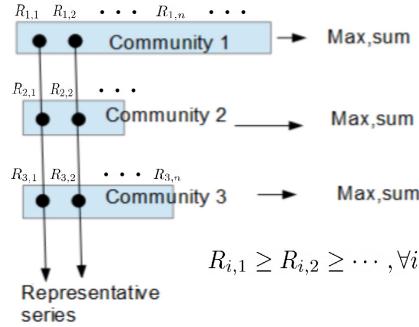


Fig. 1. The communities as the "row" sequences and the representative node series as the "column" series: the maxima and sums are taken over PRs in the communities.

3. Practice

3.1. A Study of Stationary Simulated Graphs. We consider stationary evolved graphs that are created by the PA α - $, \beta$ - and γ - schemes proposed in [6] starting from a seed network containing at least one node. α is the probability to create a new edge from a newly appearing node to an existing node and vice versa the probability γ . β is the probability to create a new edge between two existing nodes. The TI of the PR is not yet theoretically obtained. It can be only estimated. To estimate the TI we use the QQ-plot estimator proposed in [7] as the simplest and sufficiently accurate one.

In our experiment, the PA schemes with different sets of parameters (α, β, γ) were applied. The number of nodes in the evolved graphs were taken equal to $n = 10^4$. For each set (α, β, γ) , we repeatedly simulated 100 graphs. In Fig. 2, the QQ-plot estimates of the block-sums against the minimum TIs of the PRs of representative series provide a diagonal trend. Although the TIs of the block-maxima also demonstrate a trend, the latter may deviate from that for the block-sums. We omit here a detailed analysis and a stationarity testing of the representative series with the minimum TIs by the Mann-Whitney U test.

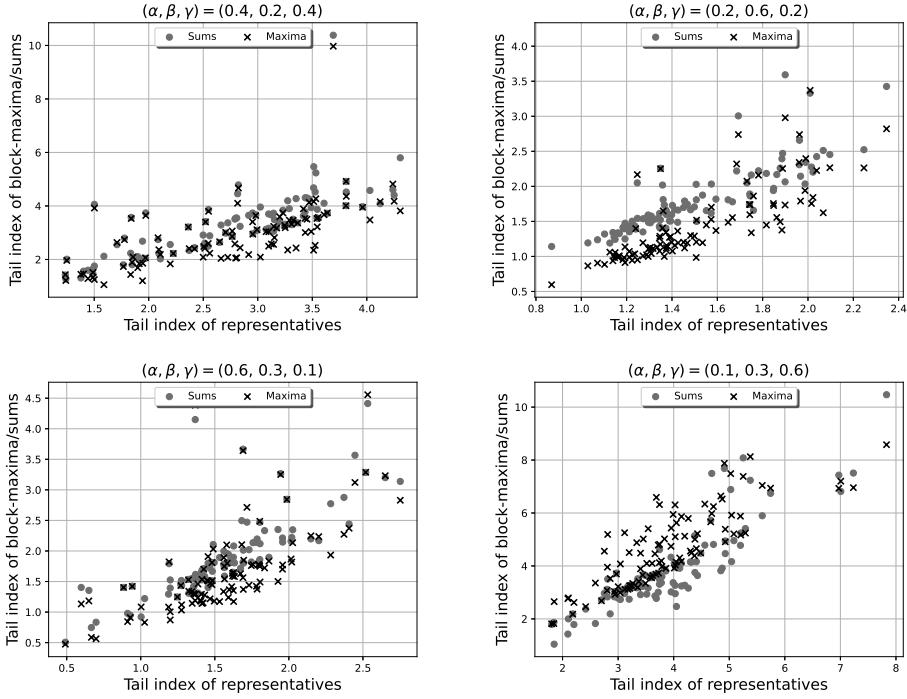


Fig. 2. The QQ plot estimates of TIs of the block-maxima and block-sums over communities against the minimum QQ plot estimates of TIs of the representative series for PRs where each point corresponds to one of the 100 graphs evolved by the PA (α, β, γ) -schemes.

3.2. A Study of Real Graphs. We analyze four temporal graphs, i.e. graphs that change with time [11]. An application of such graphs is given by gossiping and in general of information dissemination. Graphs MTH, ASK, SPR and WIKI taken from [12] were studied. Since the graphs may be huge we use a set of their subgraphs. Namely, each graph is divided into communities by the Directed Louvain's algorithm [4] and the largest communities with the number of nodes more than 1000 are used further. The latter communities are again divided into communities and their block-maxima and sums were calculated. The number of the communities is quite different which may worsen the accuracy of the TI estimation for the block-maxima and sums if the number is not large enough as for the MTH graph. We compare the TIs of the sums and maxima of the PRs over communities with the minimum TIs among the representative series. One may conclude that the block-maxima and sums over all communities in Fig. 3 have similar values of the TIs as the minimum

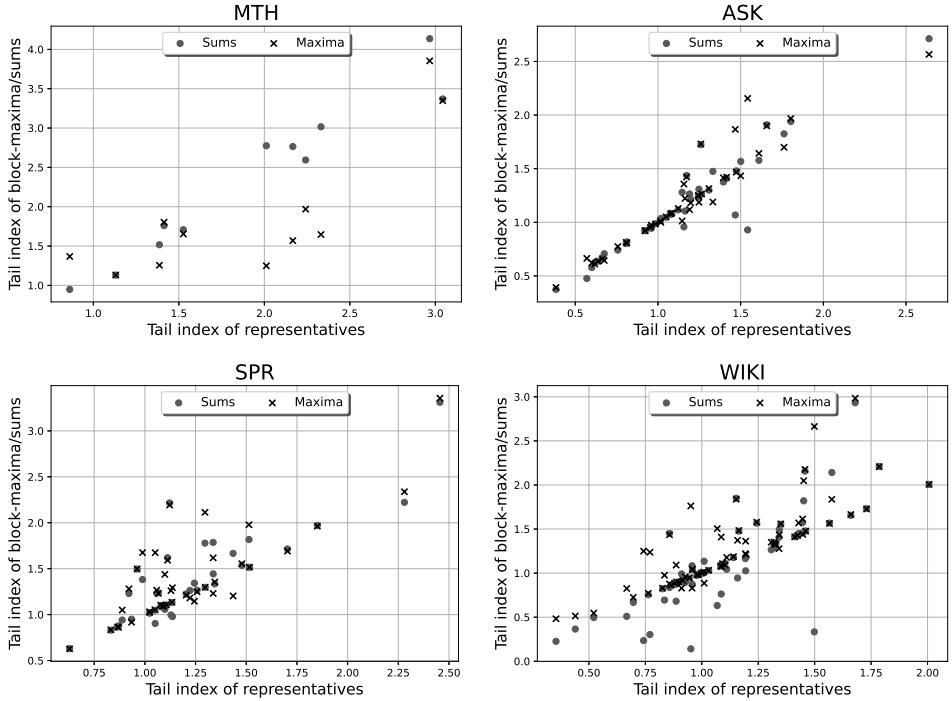


Fig. 3. The QQ plot estimates of TIs of the block-maxima and block-sums over communities against the minimum QQ plot estimates of the representative series for PRs where each point corresponds to one of the selected subgraphs of temporal graphs.

TIs among the representative series. This result is in agreement with results in [8, 9] that we aimed to check.

4. Conclusion

The prediction of TIs of the PRs and MLMs of superstar nodes in random graphs is proposed. To this end, the graph is divided into communities. The superstar node within a community is assumed to have in-coming links from all nodes of the community. The TIs of the PRs and MLMs of the superstars may be approximated by the minimum PR's TI among series constructed by the representative nodes of the communities. The obtained results confirm the theory derived in [8, 9]. The i th representative series is chosen by the i th PR maxima within each community. Since communities are of a random size, some nodes may fall in several representative series that leads to their dependence. The latter does not contradict the assumptions in [8, 9]. A random number of the representative series may have a minimum TI.

The advantage of such approach is that the communities constitute weak connected subgraphs and hence, the r.v.s of the representative series are weakly dependent that is mostly required for the TI estimation. By the theory [8, 9] the representative series with a minimum TIs have to be stationary distributed with regularly varying tails. The r.v.s of the rest of representative series may be distributed with different TIs larger than the minimum one.

Among estimators of the TI, the QQ plot estimator seems to be the most trustable metric by our study of the simulated evolving networks. More detailed analysis will be provided in the extended version.

REFERENCES

1. Volkovich Y. V., Litvak N. Asymptotic analysis for personalized web search // Adv. Appl. Prob. 2010. V. 42(2), P. 577–604.
2. Jelenkovic P. R., Olvera-Cravioto M. Information ranking and power laws on trees // Adv. Appl. Prob. 2010. V. 42(4), P. 1057–1093.
3. Markovich N.M., Ryzhov M. and Krieger, U.R. Nonparametric analysis of extremes on web graphs: pagerank versus max-linear model // Communications in Computer and Information Science. 2017. V. 700, P. 13–26.
4. Dugué N., Perez A. Directed Louvain : maximizing modularity in directed networks // [Research Report] Université d'Orléans. hal-01231784. 2015.
5. Fortunato, S. Community detection in graphs // Physics Reports. 2010. V. 486(3), P. 75–174.
6. Wan P., Wang T., Davis R. A. and Resnick S.I. Are extreme value estimation methods useful for network data? // Extremes. 2020. V. 23, P. 171–195.
7. Das B., Resnick S.I. QQ Plots, Random Sets and Data from a Heavy Tailed Distribution // Stochastic Models. 2008. V. 24(1), P. 103–132.
8. Markovich N., Rodionov I. Maxima and sums of non-stationary random length sequences // Extremes. 2020. V. 23(9), P. 451–464.
9. Markovich N. Extremes of Sums and Maxima with Application to Random Networks // 2021. arXiv:math.PR/2110.04120
10. Brin S., Page L. The anatomy of a large-scale hypertextual Web search engine // Computer Networks and ISDN Systems. 1998. V. 30, P. 107–117.
11. Michail, O. An Introduction to Temporal Graphs: An Algorithmic Perspective // In: Zaroliagis, C., Pantziou, G., Kontogiannis, S. (eds) Algorithms, Probability, Networks, and Games. Lecture Notes in Computer Science, V. 9295. P. 308–343, Springer, Cham, 2015.
12. Leskovec J., Krevl A. SNAP Datasets: Stanford Large Network Dataset Collection, 2014.<http://snap.stanford.edu/data>

УДК: 519.218

Определение среднего времени работы до отказа беспроводной сенсорной сети

О.А. Ткачев¹

¹Московский авиационный институт, Волоколамское шоссе д.4, Москва, Россия

tkachev_oleg55@mail.ru

Аннотация

Предлагается аналитическое выражение для определения среднего времени работы до отказа беспроводной сенсорной сети. Предполагается, что узлы, являющиеся сенсорами, являются невосстанавливаемыми, идентичны по надежности, отказывают независимо друг от друга и имеют экспоненциальное распределение времени работы до отказа, а узел, являющийся стоком, абсолютно надежен.

Ключевые слова: беспроводная сенсорная сеть, анализ надежности, среднее время работы до отказа, метод Монте-Карло.

1. Введение

Беспроводные сенсорные сети (БСС) являются современным и эффективным средством реализации различных систем мониторинга и управления. БСС состоят из миниатюрных вычислительно-коммуникационных устройств — сенсоров. Данные, получаемые от сенсоров, должны передаваться в один или несколько центральных узлов, называемых стоками. Как правило, БСС имеет избыточное количество сенсоров, и система является работоспособной, если определенная часть сенсоров способна устанавливать соединение со стоком. Прекращение передачи данных от сенсора к стоку возможно по двум причинам. Первая причина — отказ сенсора, вторая причина — работоспособный сенсор не может передать данные стоку, так как вследствие отказа других сенсоров разорваны все возможные пути передачи данных. В большинстве случаев критерием работоспособности сети является связность ее узлов. Сеть считается работоспособной, если связаны все узлы или определенная часть узлов. В качестве показателя надежности сети можно использовать вероятность связности заданного подмножества узлов, среднюю вероятность связности пар узлов [1, 2, 3, 4]. В более сложных случаях рассматриваются параметрические отказы [5, 6, 7]. В данной работе будет использован следующий критерий работоспособности БСС: сеть работоспособна,

если заданная часть сенсоров способна передавать данные стоку. Этот критерий работоспособности БСС был предложен в работе [8]. В значительной части работ, посвященных анализу надежности сетей, предполагается, что узлы сети абсолютно надежны, а отказывают только ее ребра. Для БСС характерен отказ узлов, а не каналов передачи данных [9], поэтому в данной работе рассматривается случай абсолютно надежных ребер и ненадежных узлов. В качестве показателя надежности используется среднее время работы до отказа. Отказом считается переход в состояние, при котором число работоспособных сенсоров, имеющих связь со стоком, меньше заданного значения.

2. Модель надежности сети, состоящей из идентичных невосстанавливаемых узлов

Структура БСС задается неориентированным графом $G(X,E)$, где: X множество вершин, а E множество ребер. Вершины графа соответствуют узлам сети (сенсорам), а ребра – каналам связи. Предполагается, что узлы, являющиеся сенсорами, невосстанавливаемые, идентичны по надежности, отказывают независимо друг от друга и имеют экспоненциальное распределение времени работы до отказа, а узел, являющийся стоком, абсолютно надежен. Вершины распределены по некоторой территории, и ребра существуют между теми вершинами, которые находятся в зоне уверенного приема радиосигнала. Обозначим Z_i – вероятности отказа сети при отказе i элементов. Значение Z_i может быть определено из выражения:

$$Z_i = \frac{Y_i}{\binom{n}{i}}.$$

где Y_i - число неработоспособных состояний сети при различных комбинациях из i отказавших элементов. Для сетей средней и большой размерности ($n > 30$) значения Z_i могут быть получены только в результате статистического моделирования, известного под названием метод Монте-Карло [10]. Вместо точного значения Z_i используется статистическая оценка . Рассмотрим непрерывный Марковский процесс (рис.1). Состояния процесса задаются числом отказавших элементов i и состоянием сети. Все неработоспособные состояния заменены одним поглощающим состоянием. Обозначим:

λ – интенсивность отказа элементов;

Z_i^* – вероятность отказа сети при отказе i -го элемента, если при наличии $i-1$ отказавших элементов сеть была работоспособна;

$\lambda''_i = (n-i)\lambda Z_{i+1}^*$ – интенсивность отказов в состоянии i ;

$\lambda'_i = (n-i)\lambda(1-Z_{i+1}^*)$ – интенсивность переходов в из работоспособного состояния i в работоспособное состояние $i+1$;

k – максимальное число элементов, после отказа которых сеть может быть работоспособна.

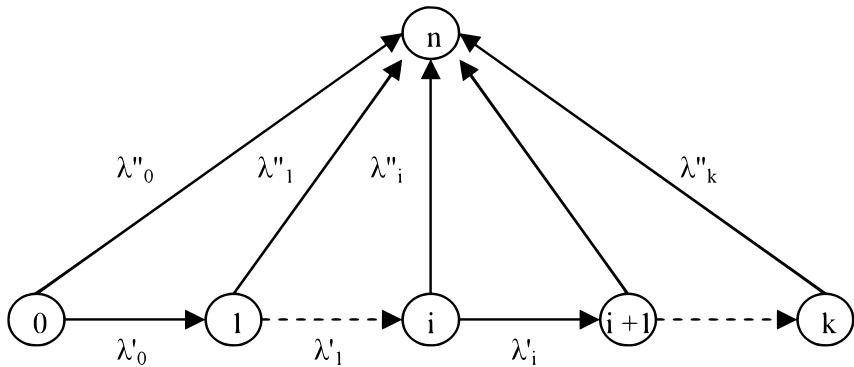


Рис. 1. Граф переходов марковского процесса изменения состояний сети

Взаимосвязь между значениями Z_i^* , и Z_i была установлена в работе [11].

$$Z_i^* = \frac{Z_i - Z_{i-1}}{1 - Z_{i-1}}. \quad (1)$$

$$\prod_{j=1}^i (1 - Z_j^*) = 1 - Z_i. \quad (2)$$

Вероятности пребывания процесса в различных состояниях $P_i(t)$ могут быть получены в результате решения системы дифференциальных уравнений:

$$\begin{cases} \frac{\partial P_0(t)}{\partial t} = -\lambda'_0 P_0(t) - \lambda''_0 P_0(t) = -\lambda_0 P_0(t); \\ \frac{\partial P_i(t)}{\partial t} = \lambda'_{i-1} P_{i-1}(t) - \lambda'_i P_i(t) - \lambda''_i P_i(t) = \lambda'_{i-1} P_{i-1}(t) - \lambda_i P_i(t); \\ \frac{\partial P_n(t)}{\partial t} = \sum_{i=0}^k \lambda''_i P_i(t); \end{cases} \quad (3)$$

Для решения системы дифференциальных уравнений (3) используем преобразование Лапласа, при начальных условиях $P_0(0) = 1, P_i(0) = 0$. Введем обозначение

$$F_i(s) = \int_0^\infty P_i(t) e^{-st} dt. \quad (4)$$

Система дифференциальных уравнений (3) приводится к системе алгебраических уравнений относительно $F_i(s)$. В результате решения системы алгебраических уравнений получим:

$$\begin{aligned} F_0(s) &= \frac{1}{s + \lambda_0}; \\ F_i(s) &= \frac{\lambda'_{i-1}}{s + \lambda_i} F_{i-1}(s); \\ F_n(s) &= \frac{1}{s} \sum_{i=0}^k \lambda''_i F_i(s). \end{aligned} \quad (5)$$

Выражение для среднего времени работы до отказа можно получить, используя свойство интегрирования преобразования Лапласа.

$$T = \sum_{i=0}^k F_i(s)|_{s=0}. \quad (6)$$

Подставляя в (5) значение $s = 0$ получим

$$F_i(0) = \frac{\lambda'_{i-1}}{\lambda_i} F_{i-1}(0) = \frac{\prod_{j=1}^i (1 - Z_j^*)}{(n-i)\lambda}. \quad (7)$$

Используя (2), выражение (7) можно упростить

$$F_i(0) = \frac{\prod_{j=1}^i (1 - Z_j^*)}{(n-i)\lambda} = \frac{1 - Z_i}{(n-i)\lambda}.$$

Тогда из (6) следует

$$T = \frac{1}{\lambda} \sum_{i=0}^k \frac{(1 - Z_i)}{n-i}. \quad (8)$$

3. Использование статистического моделирования для оценки значений Z_i

Из выражения (8) следует, что вычисление среднего времени работы до отказа сводится к определению значений Z_i . БСС являются высоконадежными системами, и отказы таких систем являются редкими событиями. Актуальной задачей является определение количества испытаний, которые необходимо провести

для получения приемлемой точности статистической оценки. Для вычисления статистической оценки \bar{Z}_i генерировались N случайных состояний сети с i отказавшими вершинами. Для каждого состояния с помощью волнового алгоритма определялось количество вершин, связанных со стоком. Состояниями отказа считались состояния, в которых число связанных со стоком вершин было меньше m . Обозначим Y_i – число состояний отказа, полученных в результате анализа N случайных состояний сети с i отказавшими вершинами. Статистическая оценка Z_i вычисляется по формуле

$$\bar{Z}_i = \frac{Y_i}{N}.$$

Среднеквадратическое отклонение σ равно

$$\sigma = \sqrt{\frac{Z_i(1 - Z_i)}{N}}.$$

Погрешность статистической оценки равна 3σ . Относительная погрешность статистической оценки

$$\varepsilon = \left| \frac{\bar{Z}_i - Z_i}{Z_i} \right| = 3 \sqrt{\frac{1 - Z_i}{N * Z_i}} \approx 3 \sqrt{\frac{1}{N * Z_i}}. \quad (9)$$

Выражение (9) позволяет определить количество испытаний, которые необходимо осуществить для определения значений Z_i , с заданной относительной погрешностью ε .

$$N = \frac{9}{\varepsilon^2 Z_i}.$$

При проведении эксперимента использовалась модель сети в виде решетки размером $9*9$, содержащая 81 узел: 80 сенсоров и один сток, который располагался в центре решетки. Для оценки абсолютной $\Delta Z_i = \bar{Z}_i - Z_i$ и относительной погрешности $\varepsilon^* = \frac{\Delta Z_i}{Z_i}$, путем полного перебора, были определены точные оценки Z_3 и Z_4 для значения $m = 75$. В результате анализа данных, полученных в результате проведения эксперимента, установлено что абсолютная погрешность $\Delta Z_i \approx 1.5\sigma$, а относительная погрешность ε^* находилась в диапазоне от 0.005 до 0.07, при $N = 10^7$. Время моделирования при $N = 10^7$ составляло $\approx 1\text{мин. } 30\text{ сек.}$, процессор Intel Core i3 -4130, 3.40GHz.

4. Заключение

Получено аналитическое выражение для определения среднего времени работы до отказа беспроводной сенсорной сети. Вычисление значения этого показателя сводится к вычислению вероятностей отказа сети при отказе определенного

количества ее узлов. Для вычисления значений этих вероятностей используется метод Монте-Карло. Выполнено аналитическое и экспериментальное исследование погрешности статистической оценки.

Литература

1. Родионов А.С., Родионова О.К., Кумулятивные оценки средней вероятности связности пары вершин случайного графа // Проблемы информатики. 2013. №19. С.3-12.
2. Синегубова С.В., Синегубов С.В. Оценка надежности систем большой размерности // Охрана, безопасность, связь. 2020. № 5-2. С. 96-100.
3. Ткачев О.А. Анализ надежности сетей, состоящих из идентичных элементов. // Надежность, №1(48) 2014, С.30-44.
4. Ткачев О.А. Анализ надежности сетей передачи данных // Труды международной конференции «Распределенные компьютерные и телекоммуникационные сети: теория и приложения» (DCCN-2010) , Москва, 2010, С.276-283.
5. Вишневский В.В., Мухтаров А.А., Першин О.Ю. Задача оптимального размещения базовых станций широкополосной сети для контроля линейной территории при ограничении на величину межконцевой задержки // Труды международной конференции «Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь» (DCCN-2020) , Москва, 2020, С.148-155.
6. Головинов Е.Э., Аминьев Д.А., Козырев Д.В., Ларионов А.А. Модель надёжности коммуникационной сети метеостанций минимальной конфигурации // Труды международной конференции «Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь» (DCCN-2019) , Москва, 2019, С.484-491.
7. Ткачев О.А. Анализ надежности вычислительных сетей с использованием параметрических критериев отказа // Экономика, статистика, информатика. Вестник УМО. № 1 2016 С.83-87.
8. Мигов Д.А. Показатель надежности для беспроводных самоорганизующихся сетей // Вестник СибГУ-ТИ. 2014. № 3. С. 3–12.
9. Мигов Д.А. Об одном показателе надежности для сетей с отказами узлов. // Проблемы информатики. 2013. №2. С.43-48.
10. Cancela H., Rubino G. Markovian models for dependability analysis Rare Event Simulation using Monte Carlo Methods Edited by Gerardo Rubino and Bruno Tuffin, 2009 John Wiley & Sons, Ltd. p. 125-144.
11. Tkachev O. A. Application of Markov chains for the reliability analysis of systems with a complex structure. Cybernetics and Systems Analysis, Volume 19, 709-716 (1983).

UDC: 519.873

Reliability of a Load-Sharing k -out-of- n System Under Decreasing of Components Residual Lifetime

Rykov V.^{1,2,3} and Ivanova N.^{1,4}

¹Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St., Moscow, 117198, Russian Federation

²Gubkin Russian State Oil and Gas University, 65 Leninsky Prospekt, Moscow, 119991, Russia

³Institute for Transmission Information Problems (named after A.A. Kharkevich)
RAS, Bolshoy Karetny, 19, GSP-4, Moscow, Russia

⁴V.A.Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65 Profsoyuznaya street, Moscow, 117997, Russia

vladimir_rykov@mail.ru, nm_ivanova@bk.ru

Abstract

The reliability characteristics of a k -out-of- n system are studied in the case when a failure of one of the system's components leads to increasing the load on others, resulting in their residual lifetimes decreasing. The situation is modeled with the help of changing the components and the system hazard rate function. The reliability function and other reliability characteristics of such a system can be calculated with the proposed approach.

Keywords: a load-sharing k -out-of- n system, dependent failures, reliability characteristics, hazard rate function

1. Introduction

The best way to improve system reliability is the redundancy technique. A typical form of redundancy is a k -out-of- n configuration. A k -out-of- n ($1 \leq k \leq n$) system is a system that consists of n components. k is the number of components in the system that must fail before the entire system fails. There are many papers devoted to their investigation due to wide applications of such a system in many spheres of human activity, including engineering, telecommunication, medicine, biology etc. [1].

This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipients Vladimir Rykov, supervision and problem setting, Nika Ivanova, data curation). This paper has been partially funded by RFBR according to the research projects No.20-01-00575A (recipients Vladimir Rykov, conceptualization, and Nika Ivanova, validation) and RSF according to the research projects No.22-49-02023 (recipient Nika Ivanova, writing-review and editing)

The development of technology puts forward new problems for researchers. So, in [2] the reliability of tethered telecommunication platform investigation posed the problem of studying the reliability of a k -out-of- n system, the failure of which depends not only on the number of failed components, but also on their location in the system. For example, for a high-altitude platform based on an eight-rotor copter, stop functioning can occur due to 2 adjacent engines' failure. Such example can be modeled as a circularly consecutive 2-out-of-8 system. A consecutive k -out-of- n system is a system with n linearly (circularly) connected components, and it fails if and only if at least k consecutive components in the system fail [3], [4].

In the real practice, component failure can be the cause of both system failure and an increase or redistribution of the load on the components remaining in operation. In the literature, the second case is defined as a load-sharing k -out-of- n system and can be interpreted differently [5]. The failure of any components can lead to an increase in the load on the remaining ones and, consequently, to a decrease in their residual lifetime. In this paper, we study the main reliability characteristics of k -out-of- n systems in conditions when the failure of one of the system components leads to a redistribution of the load on the remaining components and, consequently, to a proportional change in their residual lifetime.

2. Problem Setting and Notation

Consider a k -out-of- n system, which consists of n components and fails if and only if at least k of them fail. At that, failure of one of the components leads to increasing of the load to all other, that results decreasing their residual lifetimes. Denote by A_i ($i = 1, 2, \dots, n$) lifetimes of the system components that are supposed to be independent identically distributed (iid) random variables (rv's) and by $A(t) = \mathbf{P}\{A_i \leq t\}$ their common cumulative distribution function (cdf).

We will model the components and the whole system reliability in terms of their hazard rate functions. It is well-known that the reliability function of a component $r(t) = 1 - A(t)$ connected with its hazard rate function $\alpha(t)$ by the following expression,

$$\alpha(t) = \frac{a(t)}{1 - A(t)},$$

where $a(t)$ is a probability density function (pdf) of rv's A_i . In terms of hazard rate function, the cdf $A(t)$ of the component lifetime has a form

$$A(t) = 1 - \exp \left\{ - \int_0^t \alpha(u) du \right\}.$$

It is supposed that after any component's failure, the load on all other components increases, which leads to a decrease in their residual lifetimes. Suppose that after the

i -th failure (failure of the i -th component), all surviving components “age” for some time c_i , which corresponds to a jump in the components’ hazard rate function with the value of its shift to the time c_i taking place. In terms of the system components’ hazard rate function, this means that on the semi-interval $[S_i, S_{i+1})$ between the i -th and $(i+1)$ -th failures, the components’ hazard rate function has the form

$$\alpha_i(u) = \alpha_{i-1}(c_i + u). \quad (1)$$

where the value c_i defines the time for which all surviving components “age” after the i -th component failure.

For the system reliability study, denote by j the system state, in which j components are in failure state, and by $E = \{0, 1, 2, \dots, k\}$ the system state set. It is supposed that in the initial time all components are in UP states, that means that the initial system state is $j = 0$.

The problem under study consists in calculation of main system reliability characteristics such as reliability function and its time indicators.

3. Main Results

It is known that for any component with a hazard rate function $\alpha_0(u) = \alpha(u)$ its reliability function equals to

$$r_1(t) = \exp \left\{ - \int_0^t \alpha_0(u) du \right\}.$$

For k -out-of- n system with such components, hazard rate function is $\lambda_0(u) = n\alpha_0(u)$ (suppose that $\lambda_i(u) = (n-i)\alpha_i(u)$), and therefore its reliability function up to the first failure S_1 equals

$$R_1(t) = \mathbf{P}\{\min [A_i : i = \overline{1, n}] > t\} = \exp \left\{ - \int_0^t \lambda_0(u) du \right\}. \quad (2)$$

At that according to the rule (1) hazard rate functions of each of the surviving components are shifted by constant c_1 . Thus, for $t > S_1$ they are equal to $\alpha_1(t) = \alpha_0(c_1 + t)$ for $t > S_1$.

In order to calculate system reliability function after each component failure denote by

- $I_0 = \{1, 2, \dots, n\}$ the initial set of components,
- $S_1 = \min\{A_i : i \in I_0\}$ the time of the first failure, $i_1 = \arg \min\{A_i : i \in I_0\}$ the number of the first failed component, $I_1 = I_0 \setminus \{i_1\}$;

- analogously define by induction S_l the time moment of l -th failure, i_l the number of failed component in this time, and $I_l = I_{l-1} \setminus \{i_l\}$ the set of remaining after this step,

$$S_l = \min\{A_i : i \in I_{l-1}\}, \quad i_l = \arg \min\{A_i : i \in I_{l-1}\}, \quad I_l = I_{l-1} \setminus \{i_l\}.$$

In these notations, the following results hold.

Lemma 1. For any $i \in I_1$ conditional with respect to the time of the first failure, S_1 the reliability function of any surviving components equals to

$$r_2(t | S_1) = \mathbf{P}\{A_i > t | S_1\} = \exp \left\{ - \int_{S_1}^t \alpha_1(u) du \right\}.$$

Lemma 2. The lifetimes of survived after the first failure components are conditionally (with respect to the moment of the first failure S_1) independent and their joint reliability function for $i, j \in I_1$ are determined as follows

$$\mathbf{P}\{A_i > x, A_j > y | S_1\} = \exp \left\{ - \int_{S_1}^x \alpha_1(u) du \right\} \times \exp \left\{ - \int_{S_1}^y \alpha_1(v) dv \right\}.$$

Lemma 3. The conditional (with respect to the first failure time) reliability function of the second failure time equals to

$$R_2(t | S_1) = \mathbf{P}\{S_2 > t | S_1\} = \mathbf{P}\{\min[A_i : i \in I_1] > t | S_1\} = \exp \left\{ - \int_{S_1}^t \lambda_1(u) du \right\}.$$

Similar to the properties of survivors after the first failure, they are also performed after each of the subsequent ones.

- Lemma 4.* **1.** For any $i \in I_l$ conditional with respect to the time S_l of l -th failure, the reliability function of any surviving components equals to

$$r_{l+1}(t | S_l) = \mathbf{P}\{A_i > t | S_l\} = \exp \left\{ - \int_{S_l}^t \alpha_l(u) du \right\}.$$

- 2.** The lifetimes of survived after the l -th failure S_l components are conditional (with respect to this moment of time) and independent and for any $i, j \in I_l$ their joint reliability function is determined as follows

$$\mathbf{P}\{A_i > x, A_j > y | S_l\} = \exp \left\{ - \int_{S_l}^x \alpha_l(u) du \right\} \times \exp \left\{ - \int_{S_l}^y \alpha_l(v) dv \right\}.$$

- 3.** The conditional (with respect to the l -th failure time S_l) reliability function of a system with surviving components over the domain $t \leq S_l$ equals to

$$R_{l+1}(t | S_l) = \mathbf{P}\{S_{l+1} > t | S_l\} = \exp \left\{ - \int_{S_l}^t \lambda_l(u) du \right\}. \quad (3)$$

The above results deal with some component conditional failure time (with respect to the previous one). Based on these results, the next theorem considers the appropriate unconditional reliability functions.

Theorem 1. Reliability functions up to the l -th failure ($l = \overline{1, k - 1}$) are recursively determined by the following relations,

$$R_{l+1}(t) = - \int_0^t R_{l+1}(t|x) dR_l(x) + R_l(t), \quad (4)$$

where for the recursion beginning the function $R_1(x)$ is determined from (2) and for its continuation the function $R_l(t|x)$ is determined from (3),

$$R_1(t) = \exp \left\{ - \int_0^t \lambda_0(u) du \right\}, \quad R_{l+1}(t|x) = \exp \left\{ - \int_x^t \lambda_l(u) du \right\}.$$

Remark 1. To explain the formula (4) note that the conditional probability $R_{l+1}(t|S_l)$ is defined only over the domain $S_l \leq t$. Thus, to calculate the unconditional one, we should add except of the first term in (4) also the second one that represents the probability of the complement set $|S_l > t|$.

4. An Example: a 3-out-of-6 system

Consider a 3-out-of-6 system. Suppose that $A(t)$ is Gnedenko-Weibull (GW) distribution with mean time $a = 1$. Thus, its hazard rate function has the form

$$\alpha(t) = \theta (\Gamma(1 + \theta^{-1}))^\theta t^{\theta-1}.$$

It is well known that the behavior of this function for GW distribution depends on its coefficient of variation v . If $v = 1$, GW distribution turns to the exponential one with meantime a . The coefficient of variation $0 < v < 1$ leads to increasing of the hazard rate function $\alpha(t)$, and $v > 1$ leads to decreasing of $\alpha(t)$. Thus, the coefficient of variation v is used according to the hazard rate function is monotonic and increases. Thereby, the shape parameter of GW distribution is $\theta = 1, 2, 3$, which leads to the coefficient of variation $v = 1, 0.5227, 0.3634$.

Figure 1 presents the reliability function $R(t)$ with the 'aging' time $c_i = 0.1$, $i = 1, 2$. According to this graph, the higher reliability coincides to the lower value v .

5. Conclusion

In the paper, a k -out-of- n system is considered, in which the failure of any components leads to increasing the load to all others that leads to decreasing of their residual lifetimes. The components and the whole system reliability are modelled

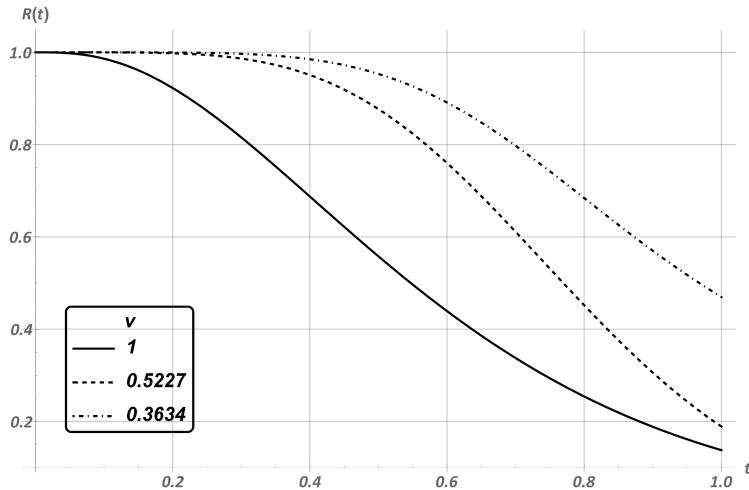


Fig. 1. Reliability function $R(t)$

in terms of their hazard rate functions. After any component failure, all surviving components “age” for some time c_i , depending on the number of the rest components.

The paper presents the approach for reliability function calculation, the implementation of which is shown in the example of 3-out-of-6 system with Gnedenko-Weibull distribution of its lifetime.

REFERENCES

- Trivedi, K.S. *Probability and Statistics with Reliability, Queuing and Computer Science Applications*, 2nd ed.; John Wiley & Sons, USA, **2016**.
- Vishnevsky, V.M.; Kozyrev, D.V.; Rykov, V.V.; Nguyen, D.P. Reliability modeling of an unmanned high-altitude mobile of a tethered telecommunication platform. *Information Technologies and Computer Systems* **2020**, 4, 26–36, doi: 10.14357/20718632200403.
- Kuo, W.; Zuo, M.J. *Optimal Reliability Modeling: Principles and Applications*, John Wiley & Sons: New York, **2003**.
- Gökdere, G.; Ng, H.K.T. Time-dependent reliability analysis for repairable consecutive- k -out-of- n :F system. *Statistical Theory and Related Fields* **2021**, doi: 10.1080/24754269.2021.1971489.
- Sutar, S.; Naik-Nimbalkar, U.V. A load share model for non-identical components of a k -out-of- m system. *Applied Mathematical Modelling* **2019**, 72, 486–498, doi: 10.1016/j.apm.2019.03.025.

UDC: 004.94

The Timeliness of The Reserved Service in the Cluster with the Regulation of the Time of Destruction of Overdue Requests in the Node Queues

V.A. Bogatyrev ^{1,3}, A.V. Bogatyrev ², S.V. Bogatyrev ^{2,3}

¹Department Information Systems Security, Saint-Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia

²Yadro Cloud Storage Development Center, Saint Petersburg, Russia

³ITMO University, Saint Petersburg, Russia

Abstract

Computer clusters operating in real time are considered. Cluster nodes are represented as single-channel service systems with limited queues. The possibilities of ensuring timely maintenance of a flow that is heterogeneous in terms of the allowable waiting time as a result of query replication and managing the number of waiting places in node queues are investigated. It is shown that with heterogeneity of flows, there are additional tasks of finding a compromise to ensure timely servicing of various flows, taking into account the impact of waiting places on the loss and timeliness of servicing requests. The ongoing research is related to the contradictory effect of increasing the places in the node queue on reducing the probability of losing requests due to buffer overflow and on increasing the waiting time for requests. Redundant service leads to an increase in the intensity of the total flow and, as a result, to an increase in the probability of losing requests, at the same time it leads to an increase in the reliability of service and to an increase in the probability of timely execution of at least one replica of requests.

Keywords: limited queue, replication, redundant service, timeliness, real time, heterogeneous flow computer system

1. Introduction

High performance, reliability and fault tolerance of distributed computer systems and networks [1-4] is achieved by reserving and consolidating resources [5-8], including when they are combined into clusters. Supporting the efficiency of data processing and transmission processes in distributed systems requires monitoring and monitoring of the system and its nodes. Based on the monitoring results, the system is adapted

to changes in its conditions and operating conditions. The adaptation of the system is carried out with its dynamic reconfiguration, controlled degradation, adaptive redistribution of request flows and resources.

For distributed real-time systems, in addition to ensuring structural reliability, it is important to support the reliability of functioning, provided that the computational process is timely.

The reliability and timeliness of query execution in distributed systems and networks can be improved as a result of their redundant maintenance [12-14]. Real-time redundant service is successful provided that at least one replica (copy) of the request is completed in a time less than the established limits [12- 14].

With redundant maintenance in the cluster, when requests are received, copies of them are created and sent to different nodes for maintenance. The result of executing the request is obtained from the replica served first in time. A replica is a copy of a request created in order to increase the accuracy of calculations, and possibly to ensure the timeliness of their execution. Ensuring timeliness (increasing the probability of service in an acceptable time) is possible as a result of the fact that with independent maintenance in different nodes of the cluster executing replicas of the request, the queue lengths and waiting times will be different. The independence of the service in the cluster nodes is due to the receipt of requests from different threads with different execution times and different number of copies being created.

It can be combined with traffic prioritization, node load balancing and other solutions aimed at managing the quality of service.

The efficiency and expediency of redundant query execution in systems represented by a group of queuing systems with infinite queues is shown in [12,14]. A study of the possibilities of increasing the probability of timely servicing of waiting-critical requests in a cluster as a result of query replication and managing the number of waiting places in node queues was conducted in [14]. The studies in [14] are limited to a flow that is homogeneous in terms of the allowable waiting time.

When the flows are heterogeneous in terms of the allowable waiting time, there are additional tasks of finding a compromise to ensure timely servicing of various flows, taking into account the impact of waiting places on the loss and timeliness of servicing requests.

An inhomogeneous flow in terms of the allowed waiting time for requests is understood as a stream formed as a result of combining several threads, for each of which different restrictions on the allowed waiting for requests in queues are set. Limits on the allowed waiting time for requests are possible for requests executed in real time. For some requests coming in the general flow, restrictions on the allowed waiting time in queues may not be introduced.

The compromise on ensuring timely servicing of various threads is intended to resolve the contradiction caused by the fact that an increase in the probability of timely execution of requests from one thread is associated with its decrease for others. Thus, an increase in the multiplicity of reservation requests aimed at increasing the probability of timely servicing of one of the threads (for example, with the shortest possible waiting time) leads to an unacceptable decrease in the probability of waiting for other threads. Finding a compromise is connected with the formulation of the vector optimization problem. The solution can be achieved by using scalar criteria or a criterion that has some physical meaning. Such a criterion can be determined based on the common interests of the system, taking into account the maintenance of a heterogeneous flow of requests. The article sets the task of forming such a generalizing criterion based on the total profit from the provision of services for the maintenance of the general flow, taking into account the profit and penalties from servicing requests with different allowable expectations.

The purpose of the work is to ensure timely maintenance of a flow that is heterogeneous in terms of the allowable waiting time as a result of query replication and management of the number of waiting places in node queues.

The timeliness of the reserved service of a real-time request will be determined by the probability of executing at least one of the generated copies of the request in the maximum allowable time without losing it due to errors and restrictions of waiting places in nodes.

The ongoing research is related to the contradictory effect of increasing queue space on reducing the likelihood of losing requests due to node buffer overflow and on increasing the waiting time for requests. The second contradiction is related to the fact that redundant service leads to an increase in the intensity of the total flow and, as a result, to an increase in the probability of loss of requests. At the same time, reservation of requests leads to increased reliability of service and to an increase in the probability of timely execution of at least one replica of requests.

2. The impact of waiting places on the timeliness of servicing a heterogeneous flow

Let's consider systems with a single server node, which we will present as a queuing model M/M/1 with a limited queue [15,16].

We will assume that one queue is formed for requests from all threads. Priorities for requests from different threads are not entered. We will assume that the queue length m , at which requests are lost, is set to be the same for all types of requests according to the allowed waiting time.

For systems with flow heterogeneity in terms of the allowable waiting time in a queue with limited waiting places, the efficiency of servicing the $i - th$ stream G_i

is determined by the probability that the request will be serviced, while its waiting time in the queue will be less than the allowable t_i .

$$G_i = (1 - r_m)F(t_i),$$

where r_m is the probability that the incoming request will find the queue fully occupied and will be lost.

For a single-channel CFR with m waiting places in the queue according to the well-known formula [15-17], the probability

$$r_m = \frac{1 - \rho}{1 - \rho^{m+1}} \rho^m$$

where, at the same time, Λ is the intensity of the input stream of requests, and v is the average time of their execution, is the probability that the time spent by the request in a queue that includes m waiting places is less than the time t [17],

$$F(t_i) = 1 - \sum_{i=1}^{m-1} \rho^i r_1 \sum_{j=0}^i \frac{(\mu t)^j}{j!} e^{-\mu t},$$

in this case, r_1 is the probability that the request will not be lost due to queue overflow

$$r_1 = \frac{1 - \rho}{1 - \rho^{m+1}} \rho.$$

A compromise on the timeliness of thread maintenance with two gradations of the allowable waiting time t_1 and t_2 can be achieved by using the following expressions

$$H_1 = (1 - r_m)F(t_1)F(t_2),$$

$$H_2 = (1 - r_m)(gF(t_1) + (1 - g)F(t_2)),$$

Where g is the fraction of a stream with an acceptable waiting time t_1 .

It should be noted that the first criterion is formed on the basis of the requirement of the same importance of ensuring a given probability of timely servicing of all flows. The second criterion determines the average probability of timely servicing of all types of heterogeneous flow.

The results of calculating the dependence of the H2 assessment of the timeliness of the total flow service on its intensity are shown in Fig.1. The calculation is performed at $v = 0.008s$, $t_1 = 0.1s$, $t_2 = 1s$. Curves 1-4 correspond to the number of revenge in the queue $m = 4, 5, 7, 50$ pcs. Curve 5 represents the difference D of the assessment of timeliness with the number of places in the queue $m = 4$ pcs. and $m = 50$ pcs. Figure 1 a corresponds to $g = 0.4$, and Figure 1 b corresponds to $g = 0.8$.

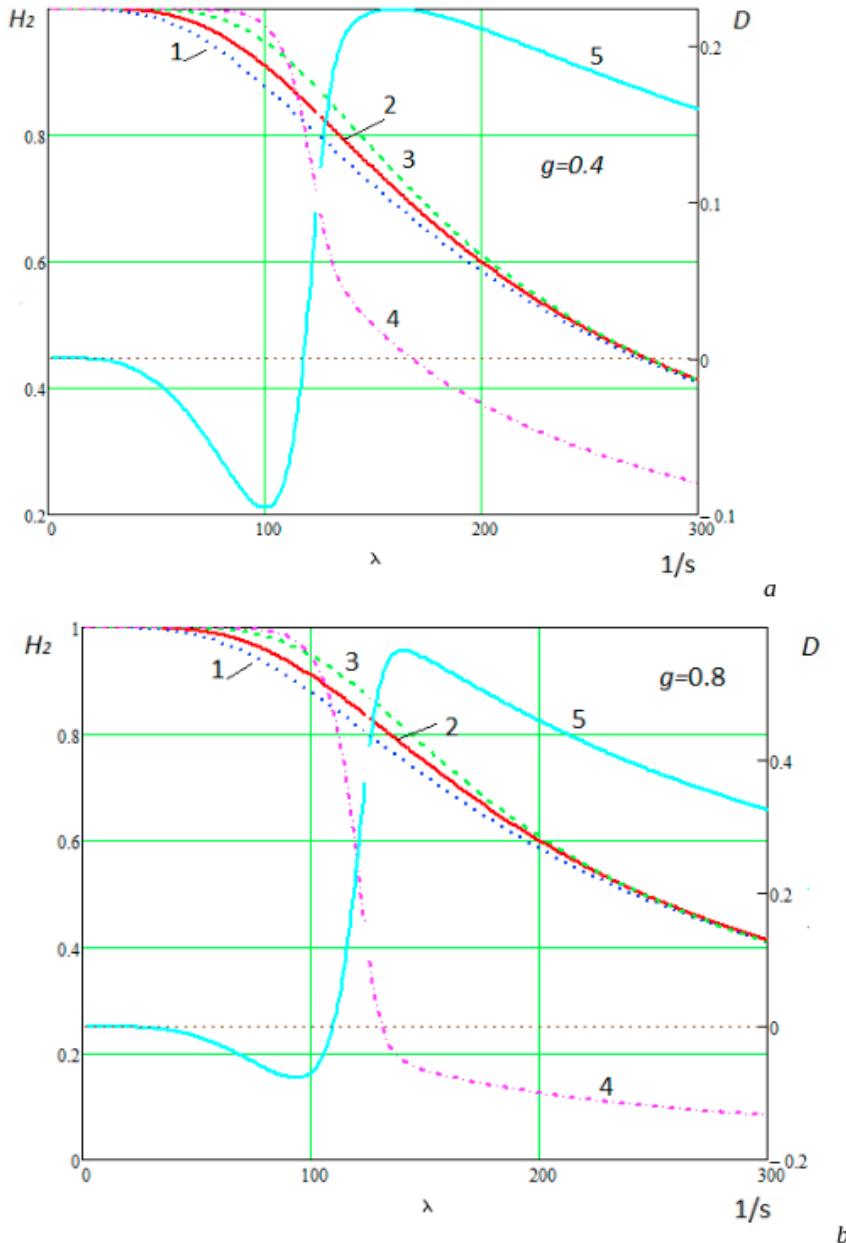


Fig. 1. The dependence of the timeliness of servicing the total flow on its intensity with a different number of waiting places in the queue.

3. The impact of reserving requests on the timeliness of servicing a heterogeneous flow in a cluster

Consider a cluster system with n nodes, each of which is represented by a CFR M/M/1 with a limited queue. For systems with heterogeneity of the flow in terms of the allowable waiting time in a queue with limited waiting places, we will analyze the effectiveness of reserved request service. Consider the case when two gradations of the allowed waiting time for requests are allocated, and only requests requiring faster service are reserved. Copies of requests are executed in different nodes of the cluster.

$$\rho = \Lambda(1 - g(d - 1))^v.$$

At the same time, Λ is the intensity of the total flow arriving at one node without increasing it due to query replication.

The probability of timely execution of at least one of the d copies of requests when reserving them is defined as

$$R(t_1) = 1 - (1 - F(t_1))^d.$$

A compromise on the timeliness of thread maintenance with two gradations of the allowable waiting time t_1 and t_2 can be achieved by using the following expressions

$$H_1 = (1 - r_m)R(t_1)F(t_2), \\ H_2 = (1 - r_m)(gR(t_1) + (1 - g)F(t_2)).$$

The first criterion reflects the requirements of the same importance of ensuring a given probability of timely servicing of all flows. The second criterion determines the average probability of timely servicing of all types of heterogeneous flow.

The assessment of the timeliness of the total flow H_1 does not allow taking into account the shares of the first and second flow. The H_2 assessment does not take into account the importance of ensuring the timeliness of the first and second streams and their impact on the overall efficiency of the system.

Both criteria do not take into account the importance of performing timely maintenance for different threads.

We formulate a generalized criterion that will allow us to take into account the impact of timely maintenance of each flow on the total economic efficiency of the system.

Let's consider an efficiency indicator that will allow us to take into account the impact of timely maintenance of each flow on the overall efficiency of the system.

We will consider the profit from timely servicing of each request of the first s_1 and the second stream s_2 and the penalties from untimely servicing of u_1 and u_2 to

be known. We will also consider the u_0 penalty for the loss of a request due to queue overflow to be set.

The efficiency indicator characterizing the intensity of income generation from the functioning of the system is defined as

$$S = \Lambda[(1 - r_m)[gR(t_1)s_1 + (1 - R(t_1))u_1 + (1 - g)F(t_2)s_2 + (1 - F(t_2))u_2] + r_m u_0].$$

The dependence of the intensity of income from the functioning of the system on the intensity of requests for the total flow of information coming to the cluster node is shown in Fig. 2, and on the number of waiting places in Fig.3. The calculation is performed at $v = 0.008$ s, $t_1 = 0.01$ s, $t_2 = 1$ s. Profit from timely servicing of requests flow $s_1 = 1$ c.u. $s_2 = 0.05$ c.u., and penalties $u_0 = -0.5$ c.u. , $u_1=-1.5$ c.u. , $u_2 = -0.05$ c.u.

In Fig.2, curves 1-3 correspond to the multiplicity of flow redundancy with the lowest allowable waiting time $d = 1, 2, 3$, the number of waiting places $m = 6$, and curves 4, 5, 6 at $m = 15$. It can be seen from the graphs that as the intensity of requests increases, it is advisable to switch the redundancy multiplicity of critical to waiting requests from $d = 3$ to $d = 2$ and then to $d = 1$.

In Fig.3, curves 1-3 correspond to the redundancy multiplicity of the stream with the lowest allowable waiting time $d = 1, 2, 3$ at the intensity of requests coming to the cluster node $\Lambda = 75$ 1/s and curves 4, 5, 6 at $\Lambda = 80$ 1/s. The graphs show the existence of an optimal number of waiting places, at which the maximum profit is achieved from servicing a flow that is heterogeneous in terms of the allowable waiting time. For the example under consideration, it is expedient to move from the redundancy multiplicity $d = 2$ to $d = 1$ as the number of waiting places increases. With an optimal value of the number of waiting places, it is advisable to reserve the flow critical to delays with a multiplicity of $d = 2$.

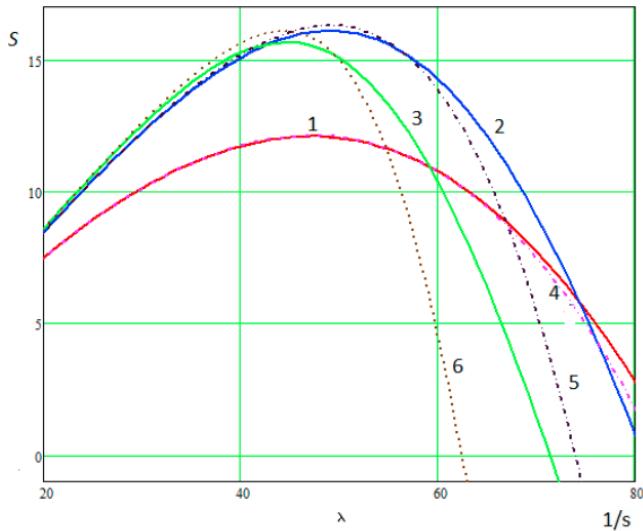


Fig. 2. Dependence of income intensity on the intensity of the total flow Λ .

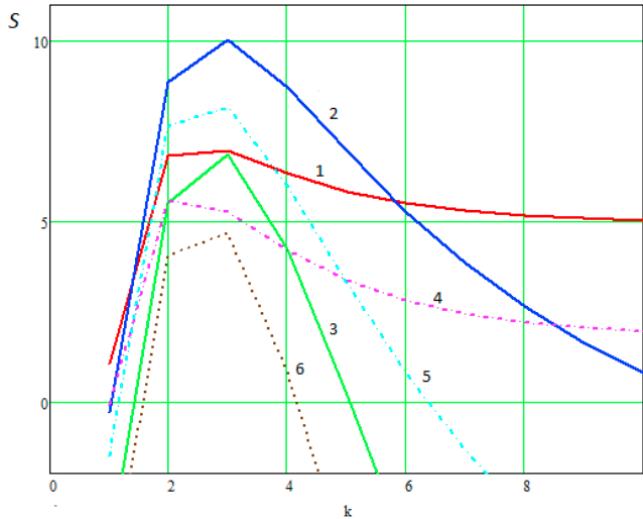


Fig. 3. Dependence of income intensity on the number of waiting places in nodes.

Thus, it is shown that, based on the regulation of the number of waiting places and the multiplicity of reservation of delay-critical requests, it is possible to ensure the reliability and timeliness of servicing a flow that is heterogeneous in terms of the allowable waiting time. managing the probability of timely error-free service of

requests as a result of regulating the multiplicity of replication and the number of waiting places allocated in node queues.

4. Conclusion

An analytical model of a cluster with the regulation of waiting places in server nodes is proposed. The possibilities of increasing the probability of timely servicing of a heterogeneous flow by acceptable delays in servicing requests with regulation of the number of places in the queues of cluster nodes and reserved servicing of the most delay-sensitive requests are shown. critical to waiting requests

REFERENCES

1. Machida F., Kawato M., Maeno Y: Redundant virtual machine placement for fault-tolerant consolidated server clusters. In: IEEE Network Operations and Management Symposium, pp. 32–39. IEEE Press, Osaka (2010), doi: 10.1109/NOMS.2010.5488431.02071-85
2. Arif Sari, Murat Akkaya (2015) Fault Tolerance Mechanisms in Distributed Systems. International Journal of Communications, Network and System Sciences, 08, 471-482. doi: 10.4236/ijcns.2015.812042
3. Merindol P. Improving Load Balancing with Multipath Routing / P. Merindol, J. Pansiot, S. Cateloin // Proc. of the 17-th International Conference on Computer Communications and Networks, IEEE ICCCN 2008. – 2008. – P. 54-61.
4. Chen, W.H. and Tsai, J.C. (2014) Fault-Tolerance Implementation in Typical Distributed Stream Processing Systems.
5. Tourouta E., Gorodnichev M., Polyantseva K., Moseva M. Providing fault tolerance of cluster computing systems based on fault-tolerant dynamic computation planning : Lecture Notes in Information Systems and Organisation. 3rd. . "Digitalization of Society, Economics and Management - A Digital Strategy Based on Post-pandemic Developments" 2022. pp. 143-150
6. Kim S., Choi Y. Constraint-aware VM placement in heterogeneous computing clusters. Cluster Comput 23, 71–85 (2020). <https://doi.org/10.1007/s10586-019-02966-6>
7. Newman M., de Castro L.A., Brown K.R. Generating fault-tolerant cluster states from crystal structures .Quantum. 2020. . 4
8. Enokido T., Takizawa M The redundant energy consumption laxity based algorithm to perform computation processes for iot services Internet of Things (Netherlands). 2020. . 9. . 100165

9. Sovetov B, Tatarnikova T, Cehanovsky V. Detection system for threats of the presence of hazardous substance in the environment. Proceedings of 2019 22nd International Conference on Soft Computing and Measurements, SCM 2019 (2019) 121-124. DOI: 10.1109/SCM.2019.8903771
10. Sahni S., Varma V. A hybrid approach to live migration of virtual machines Proc. IEEE Int. Conf. on Cloud Computing for Emerging Markets (CCEM 2012) Bengalore India pp 12–16 doi: 10.1109/CCEM.2012.6354587
11. Jin H, Li D., Wu S., Shi X,, Pan X. Live virtual machine migration with adaptive memory compression Proc. IEEE International Conference on Cluster Computing (CLUSTER '09). New Orleans, USA, 2009. Art. 5289170. doi: 10.1109/CLUSTR.2009.5289170
12. Bogatyrev V.A. , Bogatyrev A.V. , Bogatyrev, S.V. : Redundant Servicing of a Flow of Heterogeneous Requests Critical to the Total Waiting Time During the Multi-path Passage of a Sequence of Info-Communication Nodes. Lecture Notes in Computer Science// Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020. Vol. 12563. pp. 100-112. DOI 10.1007/978-3-030-66471-8_9
13. Bogatyrev V.A. , Bogatyrev A.V. , Bogatyrev, S.V, Reliability and probability of timely servicing in a cluster of heterogeneous flow of query functionality // Wave Electronics and its Application in Information and Telecommunication Systems (WECONF 2020) - 2020, pp. 9131165
14. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V. Redundant service with regulation of the number of waiting places in nodes//Wave Electronics and its Application in Information and Telecommunication Systems (WECONF 2021), 2021, pp. 9470676
15. Kleinrock L. Queueing Systems: Volume I – Theory. New York: Wiley Interscience. 1975 p. 417. ISBN 978-0471491101.
16. Kleinrock L. Queueing Systems: Volume II – Computer Applications. New York: Wiley Interscience. 1976 p. 576. ISBN 978-0471491118
17. Ovcharov L. A. Applied problems of the theory of queuing. – M. : Mechanical Engineering, 1969. – 324

UDC: 519.24

Clusters of Exceedances for Evolving Random Graphs*

Natalia M. Markovich¹ and Maksim S. Ryzhov¹

¹V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences ,
Profsoyuznaya Str. 65, 117997 Moscow , Russia

markovic@ipu.rssi.ru, nat.markovich@gmail.com,maksim.ryzhov@frtk.ru

Abstract

Evolution of random undirected graphs by the clustering attachment (CA) and with uniform node deletion is investigated. The CA causes clusters of consecutive exceedances of the modularity over a sufficiently high threshold. The modularity is a measure that allows us to divide graphs into communities. It shows the connectivity of nodes in the community. An extremal index approximates the mean cluster size and thus, it reflects a local dependence. It is shown by simulation study that estimates of the extremal index of the modularity and tail index of node degrees depend on the CA parameters.

Keywords: Random graph; evolution; modularity; clustering attachment; extremal index; tail index

1. Introduction

Network evolution attracts interest of researchers due to numerous applications. The popular mechanism to model growing network is preferential attachment (PA), see [1], [2] among others. The attachment of new nodes starts from a seed network. A newly appended node may connect to m_0 existing nodes. An existing node i is chosen randomly from the network with a probability proportional to its degree k_i : $P_{PA}(i) = k_i / \sum_j k_j$ [3]. The PA models realize a "rich-get-richer" mechanism since the earliest appended nodes get more edges. This leads to a power law node degree distribution with index α_{TI} . The distribution is determined as $P\{k_i = j\} = Cj^{-\alpha_{TI}}$, $C > 0$, $\alpha_{TI} > 0$. Another idea is the clustering attachment (CA) [3], [4]. The CA to an existing node i is provided proportional to its clustering coefficient c_i

$$P_{CA}(i) \propto c_i^\alpha + \epsilon. \quad (1)$$

*The reported study was funded by the Russian Science Foundation RSF, project number 22-21-00177 (recipient N.M. Markovich, conceptualization, methodology development, formal analysis, writing—original draft preparation; recipient M.S.Ryzhov, software, data validation).

Here $\alpha \geq 0, \epsilon \geq 0$ are attachment parameters. The clustering coefficient is

$$c_i = \frac{2\Delta_i}{k_i(k_i - 1)}, \quad (2)$$

Δ_i is the number of links between neighbours of node i or, equivalently, the random number of triangles involving node i . Since $k_i(k_i - 1)/2$ is the maximum number of triangles that may exist for node i , $c_i < 1$ holds. c_i is the probability that two random neighbours of node i are connected [4]. ϵ is mostly a dominating term in (1). It is expected in [3] that the CA does not lead to a power-law degree distribution, but to exponential tail of the degree distribution in contrast to the PA. We aim to check this hypothesis by estimation of the node tail index (TI).

Another important observation is that the CA gives rise to a specific cluster structure of the community appearance during the network growing. To be precise, the communities arise sequentially during the evolution that forms the cluster structure of the modularity Q . Let us recall that the modularity is a measure of the connectivity degree of nodes. A community consists of nodes that are strongly connected with each other and weakly connected with nodes from other communities [5]. The ability to detect community structure in a network may have practical applications. Grouping by interest in a social network or pages on related topics on the web might be represented as communities. Identification of these communities could help to understand and exploit networks more effectively.

We aim to study the clusters of consecutive exceedances of the normalized modularity $\psi(t) = Q(t)/\langle Q \rangle - 1$ over a sufficiently high threshold u , where $\langle Q \rangle$ denotes the average of the modularity values over some evolution period. The clustering of $\psi(t)$ is enhanced when α in (1) grows, see Fig.1(b) in [3]. The evolved modularity series allows us to study clustering properties of random sequences instead of graphs. Note that nodes in the graph cannot be definitely enumerated. The clusters of exceedances correspond to consecutive large values of the modularity and hence, to strongly connected communities in the network at some evolution steps. The extremal index (EI) measures a local dependence in the sense that it approximates the reciprocal of the mean cluster size [6]. The TI shows the heaviness of the distribution tail. It will be estimated by the Hill's estimator with a bootstrap selected number of the largest order statistics k [7].

Our objective is to establish an influence of parameters α and ϵ in (1) on the estimated EI and node TI. We check these relations for the case when an uniformly chosen node is deleted every time when a new node is appended. A total number of nodes remains thus fixed.

Our results are based on the simulation study. Undirected graphs are considered.

The paper is organized as follows. Required definitions and results are given in Sect. 2. Main results are provided in Sect. 3. We finalize with conclusions in Sect. 4.

2. Related works

2.1. Clustering attachment for graph evolution. The CA simulates an evolutionary growth of graphs starting from an initial seed graph $G_0 = (V_0, E_0)$. V_0 and $\|V_0\|$ are a set of vertices and their number. Similar notations E_0 and $\|E_0\|$ are applied for edges. A new node i is appended to m_0 existing nodes. $m_0 \geq 2$ may be taken as in [3]. $m_0 = 2$ means that the new node i may get Δ_i that is equal to 0 or 1 in (2). ϵ denotes a non-zero attachment probability to node i that is proportional to $P_{CA}(i)$ when $c_i = 0$ holds.

2.2. Modularity. For an undirected graph $G = (V, E)$, the modularity Q is a measure to partition the network into communities [8]. Q shows how many edges exist within communities and between them:

$$Q = \frac{1}{2e} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2e} \right] \mathbf{1}(i, j), \quad (3)$$

$e = \frac{1}{2} \sum_{ij} A_{ij}$ is a number of edges in G , A is an adjacency matrix, $\mathbf{1}(i, j)$ is equal to 1 when nodes i and j belong to the same community, k_i is a degree of node i . We use the Greedy Modularity Maximization Algorithm (GMMA). It is a hierarchical agglomeration algorithm for detecting communities which is faster than many competing algorithms: its running time on a network with n vertices and m edges is $O(md \log(n))$, where d is a depth of the dendrogram describing the community structure [8].

2.3. Extremal index. A stationary sequence $\{Y_n\}_{n \geq 1}$ of random variables (rvs) with distribution function (df) $F(x)$ and $M_n = \max_{1 \leq j \leq n} Y_j$ is said to have EI $\theta \in [0, 1]$ if for each $0 < \tau < \infty$ there is a sequence of real numbers $u_n = u_n(\tau)$ such that

$$\lim_{n \rightarrow \infty} n(1 - F(u_n)) = \tau \quad \text{and} \quad \lim_{n \rightarrow \infty} P\{M_n \leq u_n\} = e^{-\tau\theta}$$

hold [6]. The EI plays a key role in the extreme value analysis since it allows to get a limit distribution of sample maximum of rvs when the latter is dependent, i.e. $P\{M_n \leq u_n\} = F^{n\theta}\{u_n\} + o(1)$ as $n \rightarrow \infty$. The EI is equal to one for independent identically distributed rvs. The converse is incorrect. As closer θ to zero, as stronger the local dependence or clustering. The EI measures the local clustering tendency of high threshold exceedances. Its reciprocal $1/\theta$ approximates the mean number of exceedances per cluster (the mean cluster size).

To estimate θ we use the intervals estimator proposed in [9] that is one of the most effective and simple known estimators. Taking the exceedance times $1 \leq S_1 <$

$\dots < S_{N_u} \leq n$ of $\{Y_n\}_{n \geq 1}$, the observed interexceedance times are $T_i = S_{i+1} - S_i$ for $i = 1, \dots, N_u - 1$. $N_u = \sum_{i=1}^n \mathbf{1}\{Y_i > u\}$ is the number of observations exceeding a predetermined high threshold u . We denote $L \equiv L(u) = N_u - 1$. The intervals estimator is defined as

$$\hat{\theta}_n(u) = \begin{cases} \min(1, \hat{\theta}_n^1(u)), & \text{if } \max\{T_i : 1 \leq i \leq L\} \leq 2, \\ \min(1, \hat{\theta}_n^2(u)), & \text{if } \max\{T_i : 1 \leq i \leq L\} > 2, \end{cases} \quad (4)$$

where

$$\hat{\theta}_n^1(u) = \frac{2(\sum_{i=1}^L T_i)^2}{L \sum_{i=1}^L T_i^2}, \quad \hat{\theta}_n^2(u) = \frac{2(\sum_{i=1}^L (T_i - 1))^2}{L \sum_{i=1}^L (T_i - 1)(T_i - 2)}. \quad (5)$$

The intervals estimator requires a choice of u as parameter. We find u by discrepancy method proposed in [10].

3. Main results

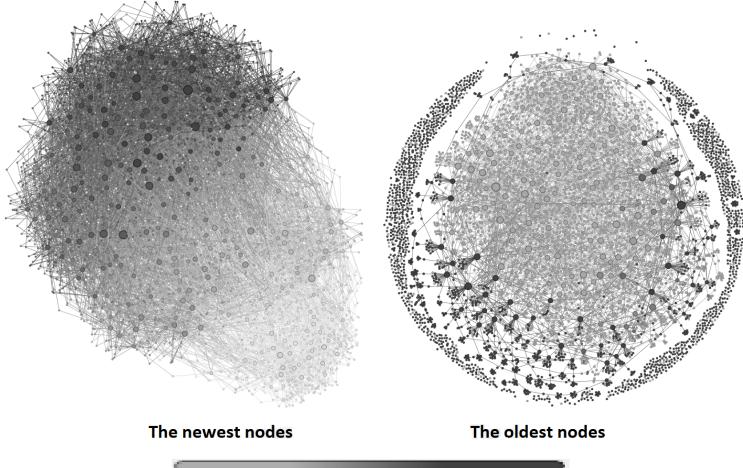


Fig. 1. A seed graph with number of nodes 5000 obtained by the CA (2), (6) with $\alpha = 1$, $\epsilon = 0$ and $m_0 = 2$ (left) and the graph evolved from the seed graph by the same CA using an uniform node deletion (right). The point size is proportional to the node degree.

Let $G(t) = (V(t), E(t))$ denote the graph at evolution step t . We consider the CA with $m_0 = 2$ throughout. The normalized measure

$$P_{CA}(i, t) = \frac{c_i^\alpha(t) + \epsilon}{\sum_{j \in V(t)} c_j^\alpha(t) + \|V(t)\| \epsilon} \quad (6)$$

is used instead of (1). Examples of a seed graph and its evolution by the CA are shown in Fig. 1. An uniform node deletion at each step of the evolution causes a large number of isolated nodes since the number of triangles decreases. In Fig.

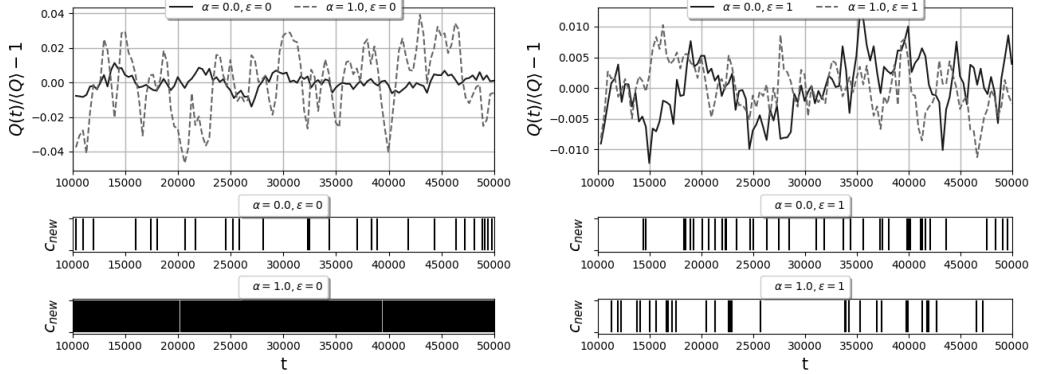


Fig. 2. The evolution of the normalized graph modularity against the evolution steps (top); and spike trains denoting injections of a new triangle when new nodes are appended and their clustering coefficient c_{new} is equal to one (two bottom lines).

2 we observe a cluster structure of time series $\psi(t)$ against t for four pairs of CA parameters (α, ϵ) . An uniform node deletion is provided. $\langle Q \rangle$ denotes the average over evolution steps in the interval $t \in [10^4, 5 \cdot 10^4]$. Spike trains in Fig. 2 indicate evolution steps that are accompanied by creation of a new triangle. It leads to increasing of the modularity and thus, to its clusters of exceedances.

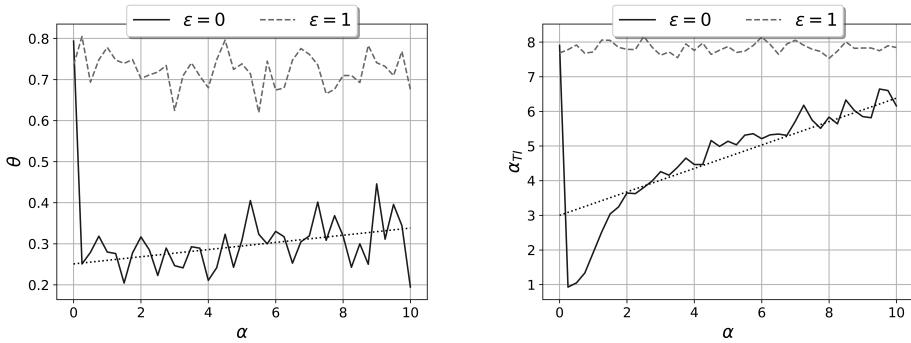


Fig. 3. Interval estimates (4), (5) of the EIs for normalized modularity $Q/\langle Q \rangle - 1$ (left) and tail indices of node degrees α_{TI} (right) averaged over 30 simulated graphs evolved by the CA with the uniform node deletion.

The case $\epsilon = 0$ is specific. If additionally node i does not belong to any triangle

of nodes and $c_i = 0$ holds not for all $i \in V(t)$, then $P_{CA}(i, t) = 0$ follows by (6). The latter implies that new nodes cannot be attached to node i . The clustering dynamic with regard to α is shown in Fig. 3. It shows that the EI for $\epsilon = 0$ tend to increase as α increases and they are closer to zero than the EI for $\epsilon > 0$ which tend to be stable. The smaller EI, the stronger clustering (or local dependence) of modularity sequences. For any positive value of ϵ , the EI is stable with regard to $\alpha > 0$. $(\alpha, \epsilon) = (0, 0)$ leads to a large EI. The TIs for $\epsilon = 0$ tend to increase as α increases as far as the TIs are stable and have larger values for $\epsilon > 0$. This means that the distribution tails become lighter as α increases for $\epsilon = 0$. $\alpha = 0$ corresponds to a constant $P_{CA}(i, t)$ for any i and any ϵ . This implies that a node i can be attached to existing nodes irrespective of the number of its triangles Δ_i .

4. Conclusion

The evolution of undirected graphs by the CA with uniform node deletion is studied. The EI of the graph modularity and the TI of node degrees are estimated. The estimates are compared by the CA parameters. By simulation, it is found that the positive values of ϵ in (1) lead to stable large values of the EI and the TI. It means a weak clustering of the modularity and a weak heaviness of node degree distribution tails. $\epsilon = 0$ causes a strong clustering since the EI that are close to zero, and the heaviness of tail of the node degree weakens since the TI α_{TI} increases as $\alpha > 0$ increases. A similar study and comparison will be provided without node and edge deletion in an extended version.

REFERENCES

1. Norros I., Reittu H.: On a conditionally poissonian graph process // *Adv. Appl. Prob.* (SGSA). 2006. V. 38, P. 59–75.
2. Wan P., Wang T., Davis R. A. and Resnick S.I. Are extreme value estimation methods useful for network data? // *Extremes*. 2020. V. 23, P. 171–195.
3. Bagrow J., Brockmann D.: Natural Emergence of Clusters and Bursts in Network Evolution // *Physical Review X*. 3. 2012. 10.1103/PhysRevX.3.021016
4. Van Der Hofstad R., Southwell A., Stegehuis C. Counting Triangles in Power-Law Uniform Random Graphs // *The Electronic Journal of Combinatorics*. 2020. V. 27(3).
5. Fortunato, S. Community detection in graphs // *Physics Reports*. 2010. V. 486(3), P. 75–174.
6. Leadbetter, M.R., Lingren, G., Rootzén, H.: *Extremes and Related Properties of Random Sequence and Processes*. ch.3, Springer, New York, 1983.

7. Markovich, N.M. *Nonparametric Analysis of Univariate Heavy-Tailed data: Research and Practice*. Chichester, West Sussex: Wiley, 2007.
8. Clauset, A., Newman, M. E. J., Moore, C. Finding community structure in very large networks // Physical Review E. 2004. 70(6), 066111. <https://doi.org/10.1103/PhysRevE.70.066111>
9. Ferro, C.A.T., Segers, J. Inference for Clusters of Extreme Values // J. R. Statist. Soc. B.. 2003. V. 65. P. 545–556.
10. Markovich N.M., Rodionov I.V.: Threshold selection for extremal index estimation // ArXiv e-prints, arXiv:2009.02318. 2020.

UDC: 004.032.26

Improvement of CNN-Based Model for Object Classification in Aero Photographs

V.T. Nguyen¹, F.F. Pashchenko^{1,2}, T.A. Bui¹, D.T. Le³

¹Moscow Institute of Physics and Technology, Moscow, Russia, 9 Institutskiy per., Dolgoprudny, Russia

²V. A. Trapeznikov Institute of Control Sciences, 65 st. Profsoyuznaya, Moscow, Russia

³Le Quy Don Technical University, Hanoi, Viet Nam

van.chong.nguen@phystech.edu

Abstract

A review of methods for solving computer vision problems by classes proved the advantages of the neural network method over all those proposed earlier. The principle of operation of neurons in an artificial network makes it possible to identify its activity, which is determined by its parameters for classifying objects in aerial photographs. The article presented a well-founded mathematical model and the main technologies of the convolutional neural network model based on the dropout technique, which is widely used in preventing network retraining. This model helps to increase the invariance to the scale of the input images. At the output of the convolutional layers of the network, several layers of the neural network are additionally installed that perform the work of the classifier.

Keywords: aerial imagery, aerial photography, aerial image classification parameters, neural network

1. Introduction

The technology of creating machines that have the ability to see is the basis of computer vision. This discipline refers to artificial systems that can perform recognition tasks. The above section provides the basic theory of object recognition, which is the basis of the integrated classification.

Difficulties with distortions in aerial images, such as image resizing and image rotation, are solved by modeling a neocognitron, which uses a qualitatively new architecture. The architecture of the neocognitron is based on the organization of the visual system of living beings. The considered basic principles of training neural networks with their features have shown that effective object recognition training

can be achieved precisely by the method of error back propagation. But if there is an obvious sign somewhere, then the result of applying the filter to this area of the image will give a large value and a small one for other parts of the image.

2. The degree of study of the problem

The results of [1, 2, 3], which show the possibility of using a neural network to approximate arbitrary functions with a given accuracy, were used as a starting point for substantiating the algorithm for neural network estimation of color channel parameters. Thus, in the work of S. Khaykin [4], a proof of the Hecht-Nielsen theorem is shown, in which the fundamental possibility of representing a continuous arbitrary function of many variables using a neural network with direct signal propagation when the network contains at least one layer of neurons is proved. Structurally, such a neural network consists of N input neurons, at least $2N+1$ hidden neurons with sigmoidal activation functions of the type and M output neurons with unknown activation functions. The results of this theorem are somewhat extended. It is proved that the parameters of the sigmoidal activation function can be set a priori, and a linear activation function of the form can be used in the output layer of neurons:

$$g(x) = ax + b \quad (1)$$

where g - linear activation function; x - the total input signal of the neuron, a, b - coefficients.

Note that the described type of neural network model with one layer of hidden neurons is called a two-layer perceptron, and with several layers of hidden neurons - convolution. It is proved that when certain structural rules are met (a sufficient number of hidden neurons), it is possible to approximate an arbitrary smooth function using a neural network with radial basis functions. Thus, convolutional neural network models are subject to theoretical verification.

Since the state of security of the information system depends on the events that occur in it, and are characterized by a set of certain controlled security parameters, the model for identifying objects in aerial photographs can be written as follows:

$$\exists s(t) \in S_a(t) \wedge p(t) \in P_a(t) \Rightarrow A \quad (2)$$

$$\exists s(t) \in S_n(t) \wedge p(t) \in P_n(t) \Rightarrow A \quad (3)$$

$s(t)$ - a lot of events that occurred in the information system; $p(t)$ - the set of values of the object classification parameters in the aerial images of the information system at the time t ; $S_n(t), P_n(t)$ - a lot of events in the information system and a lot of values of safety parameters characteristic of the implementation of signs of violation of the classification parameters of objects in aerial photographs; $S_n(t), P_n(t)$ - a lot of events

and a lot of values of security parameters characteristic of the normal state of the information system at the time t ; A - implementation of objects in aerial photographs. In addition to expressions (2, 3), expressions (4, 5) can be used to identify the normal state of the classification parameters of objects in aerial photographs:

$$\exists s(t) \in S_a(t) \wedge p(t) \in P_a(t) \Rightarrow N \quad (4)$$

$$\exists s(t) \in S_n(t) \wedge p(t) \in P_n(t) \Rightarrow N \quad (5)$$

where N - the normal state of the system.

The method of detecting objects in aerial photographs based on (5 and 6) was called "abuse detection", and the method of detecting objects in aerial photographs was called "anomaly detection" [5]. Using (6), it is possible to write a generalizing model for detecting objects in aerial photographs in the form of a continuous function of many variables:

$$\begin{cases} U = F(s(t), p(t), S_a(t), P_a(t), P_n(t)) \\ U \in (A, N) \end{cases} \quad (6)$$

It can be noted that a similar model for detecting objects in aerial images for an information system was used in [2]. At the same time, the model (6) is generalized. In many cases, only individual components are used to detect attacks. As a result, the application of the Hecht-Nielsen theorem and the results of D. Park and I. Sandberg to the functional (6) allows us to assert that with the help of a multilayer perceptron and a neural network with radial basis functions, it is possible to recognize objects in aerial photographs in information systems with a given accuracy. At the same time, a necessary condition for the verification of the neural network model is the ability to represent the classification parameters of objects in aerial photographs in the form of continuous functions.

3. Research methods

Next, we will consider the use of the obtained result on a specific example. Let's define the KDD-99 database, which contains examples of the normal functioning of the information system and the signatures of network objects in aerial photographs. It is required to show: the guarantee of recognition of the presented network objects in aerial photographs using neural network models of the form of a multilayer perceptron and a neural network with radial basis functions. The KDD-99 database contains about 5000000 records – aerial survey objects registered at certain intervals [1]. Each record consists of 42 fields. The fields from 1 to 41 contain such network connection parameters as duration, protocol type, network service, number of bytes received, number of bytes transmitted, connection status, etc. The 42 field contains

information characterizing the state of the classification parameters of objects in aerial photographs – either the absence of signs of violation of the classification parameters of objects in aerial photographs (normal), or its type. The database contains 22 types of classification features of objects in aerial photographs, which are divided into 4 main classes – recognition system failure, the presence of noise, shadow interference, color correction.

Therefore, to identify objects in aerial photographs, it is possible to use only the values of 41 object recognition parameters. This allows you to rewrite (6) as a function (7):

$$\begin{cases} U = F(p_1, p_2, \dots, p_{41}) \\ U \in (A_1, A_2, \dots, A_{22}, N_1) \end{cases} \quad (7)$$

so p_1, p_2, \dots, p_{41} - classification parameters of objects in aerial photographs; A_1, A_2, \dots, A_{22} - types of objects in aerial photographs; N_1 - the normal state of the information system.

The application of the Hecht-Nielsen theorem and the results of D. Park and I. Sandberg to function (7) indicates the possibility of using a convolutional neural network network to identify objects in aerial photographs.

To create a model for estimating the parameters used in neural network tools for recognizing objects in aerial photographs, three sets were used – a set of possible objects in aerial photographs on the PIC (**Ka**), there are many possible parameters for classifying objects in aerial photographs (**X**) and a set of output parameters of the neural network for recognizing objects in aerial photographs (**Y**). In the general case:

$$\begin{aligned} \mathbf{Ka} &= \bigcup_{j=1}^J Ka_j \\ \mathbf{X} &= \bigcup_{i=1}^I X_i \\ \mathbf{Y} &= \bigcup_{g=1}^G Y_g \end{aligned} \quad (8)$$

where Ka_j - j -th of objects in aerial photographs; J - the number of possible objects in aerial photographs; G - number of output parameters of the neural network; Y_g - g -th output parameter; X_i – i -th security parameters; I - number of security parameters.

Each element (type of objects in aerial images) of the set Ka a subset of output parameters is matched Y_N (the values of which indicate the presence of objects

in aerial photographs), consisting of elements of the set Y . In turn, each Y_N a subset of the input parameters is matched X_V (necessary to identify objects in aerial photographs), which consists of elements of the set X . Thus, in order to identify objects in aerial photographs, a set of triples "input parameters -output parameters - objects in aerial photographs" should be formed:

$$X_V \rightarrow Y_N \rightarrow Ka = \bigcup_{j=1}^J Ka_j \rightarrow X = \bigcup_{i=1}^I X_i \rightarrow Y = \bigcup_{g=1}^G = \\ (\{X_{1,1}, X_{1,2}, \dots, X_{1,V1}\} \rightarrow \{Y_{1,1}, Y_{1,2}, \dots, Y_{1,V1}\}) \rightarrow Ka_j \quad (9)$$

The use of a neural network provides that the domain of the sets X and Y are defined on $[0 \dots 1]$.

With known triples (10), the identification of objects in k -type aerial images is reduced to establishing a correspondence between the current values of the incoming parameters $\{x_{k,1}, x_{k,2}, \dots, x_{k,Y_k}\}$ and the values of the set of their parameters $\{x_{k,1}, x_{k,2}, \dots, x_{k,Y_k}\}$, which correspond to the values of the output parameters $\{Y_{k,1}, Y_{k,2}, \dots, Y_{k,Y_k}\}$, which testify about Ka_k :

$$\{x_{k,1}, x_{k,2}\} = \{x_{k,1}, x_{k,2}, \dots, x_{k,Y_k}\} \Rightarrow \{Y_{k,1}, Y_{k,2}, \dots, Y_{k,Y_k}\} \Rightarrow Ka_k \quad (10)$$

The model (10) is detailed taking into account the developed approaches to the recognition of classified and unidentified objects in aerial photographs. To do this, the set of possible objects in aerial images is presented in the following form:

$$Ka = (Ks, Kq) \quad (11)$$

where Ks, Kq - accordingly, there are a lot of classified and unidentified objects in aerial photographs.

The set of input parameters X is also divided into two parts:

$$X = (Xs, Xq) \quad (12)$$

Xs - a variety of security parameters used to recognize unidentified objects in aerial photographs, Xq - a variety of security parameters used to recognize classified objects in aerial photographs. Substituting (10, 11) into (12), we get:

$$\{xs_{k,1}, xs_{k,2}, \dots, xs_{k,Y_k}\} \cong \{Xs_{k,1}, Xs_{k,2}, \dots, Xs_{k,Y_k}\} \Rightarrow \{Ys_{k,1}, Ys_{k,2}, \dots, Ys_{k,Y_k}\} \quad (13)$$

$$\{xq_{k,1}, xq_{k,2}, \dots, xq_{k,Y_k}\} \cong \{Xq_{k,1}, Xq_{k,2}, \dots, Xq_{k,Y_k}\} \Rightarrow \{Yq_{k,1}, Yq_{k,2}, \dots, Yq_{k,Y_k}\} \quad (14)$$

The use of (13 and 14) allows us to define a model of neural network evaluation of input parameters to identify k -th unidentified objects in aerial photographs or classified objects in aerial photographs in the following form:

$$\{xs_{k,1}, xs_{k,2}, \dots, xs_{k,Y_k}\} [nnet] \{Xs_{k,1}, Xs_{k,2}, \dots, Xs_{k,Y_k}\} \rightarrow \\ \{Ys_{k,1}, Ys_{k,2}, \dots, Ys_{k,Y_k}\} \Rightarrow Ks_k \quad (15)$$

$$\{xq_{k,1}, xq_{k,2}, \dots, xq_{k,Y_k}\} [nnet] \{Xq_{k,1}, Xq_{k,2}, \dots, Xq_{k,Y_k}\} \rightarrow \\ \{Yq_{k,1}, Yq_{k,2}, \dots, Yq_{k,Y_k}\} \Rightarrow Kq_k \quad (16)$$

where $[nnet]$ - neural network comparison operator.

Given (15, 16), generalized expressions for neural network evaluation of current input parameters can be represented as:

$$xs_1[nnet]Xs_i \rightarrow Ys_i \quad (17)$$

$$xq_1[nnet]Xq_i \rightarrow Yq_i \quad (18)$$

Note that, according to the developed approach to recognizing objects in aerial photographs and results [3], in (18) it is necessary to take into account the dependencies of Xs_i and xs_i on the service life. Therefore:

$$xs_1(t)[nnet]Xs_i \rightarrow Ys_i \quad (19)$$

Thus, when identifying unidentified objects in aerial images, a neural model should be used to classify time series of data, which, according to [6], can cause significant difficulties. To overcome these difficulties, it is proposed to carry out preliminary processing of the classified parameters of the classification of objects in aerial photographs in order to remove time dependence from them. According to the results of [7], it is proposed to determine the time dependence using an additional Markov model of classified safety parameters. This allows you to modify (20) as follows:

$$xs_1(t) - \overleftarrow{X}s_1(t)[nnet]Xs_i(t) - \overleftarrow{X}s_1(t) \rightarrow Ys_i \quad (20)$$

where $\overleftarrow{X}s_1(t)$ - calculated using a convolutional model, the value of the classification parameters of objects in aerial photographs at a time t .

4. Research results

One of the tasks of developing effective neural network tools for estimating the parameters of object classification in aerial images of an information system is to determine the nomenclature of input parameters of a neural network model. The need to solve this problem is explained by the following factors:

- using as input parameters of a neural network model a large number of object classification parameters in aerial images of an information system significantly increases the amount of computing resources and complicates the process of accumulating training examples;
- the use of uninformative object classification parameters in aerial photographs leads to the training of a neural network model on noisy data, which negatively affects the correctness of the classification of unknown examples and increases the amount of computing resources;
- extraction from the input parameters of the neural network model of informative parameters of the classification of objects in aerial photographs can lead to a complete loss of its classification properties.

At the same time, studies [8, 9] indicate that the final decision on the nomenclature of input parameters of a neural network is made as a result of rather lengthy comparative experiments. To reduce the number of these experiments, it is advisable to determine the importance of each of the possible safety parameters. Since modern formalized methods for assessing the importance of the classification parameters of objects in aerial photographs do not meet the accuracy requirements [10], it was decided to use expert evaluation. It is proposed to use the method of paired comparisons, which is explained by its proven effectiveness in cases of a large number of test objects with which the classification parameters of objects in aerial photographs are associated [11].

In this case, the input data of the model is a vector, the elements of which will be matrices of expert assessments of the significance of security parameters.

The input information of the model is the entered set of characteristics of the aerial survey object (O), which is determined by the expression (21), and the output of the model is a set of safety parameters, the evaluation of which allows you to determine the gradual and unexpected objects in the aerial survey images, characteristic of the aerial survey object of the information system. The model consists of five basic processes that relate to integration processes.

The value of the classification parameters of objects in aerial photographs does not depend on time, and the realization of objects in aerial photographs (the presence of a virus) is evidenced by a certain combination of their values, which, from the point of view of statistical data approximation, is unexpected. Therefore, the detection of web-oriented viruses is classified as the detection of unidentified attacks, and when evaluating the classification parameters of objects in aerial photographs, it is taken into account that $Ka = Kq, X = Xq$. In order to reduce the volume of the analyzed neural models of input parameters, the procedure of expert evaluation of the significance of parameters specified by expressions (20) was used. As a result, a

set of input parameters is defined that correspond to the names of class comparison operators.

Two normalization techniques are popular today: local response normalization and batch-normalization. Both normalizations are used to prevent a decrease in the learning rate of network parameters. Local response normalization of compliance is an independent layer of the network. This operation normalizes each value of the input matrix by channel. Batch normalization, first introduced in the work of S. Ioffe [12], applies the standard normalization of the obtained values, and then linearly transforms them (22):

$$y = \frac{x - \mu}{\sqrt{\sigma^2 - \epsilon}} * \gamma + \beta \quad (21)$$

here ϵ, γ, β are the parameters to be configured. μ and σ - the mean and variance depend on what stage the network is at.

The question of when exactly to apply normalization (normalization layer) is still open. To perform a dropout, it does not matter that normalization is applied before or after the activation function. Given this, possible applications:

- Convolution Layer Classifier (fully connected layer) → BN (normalization) → Activation → Dropout → ...
- Convolution Layer Classifier (fully connected layer) → activation function → BN → Dropout → ...
- Convolution Layer Classifier (fully connected layer) → activation function → Dropout → BN → ...
- Convolution Layer Classifier (fully connected layer) → Dropout → BN → activation function → ...

To build a future network, it was decided to perform BN immediately after applying the convolution operation and before applying the nonlinearity operation (activation function). In addition to the maximum subsampling, these layers can perform other functions presented in Table 1, for example, an averaging subsampling or even an L2-normalized subsampling.

As a starting point for substantiating the algorithm for neural network estimation of color channel parameters, the results of Russian scientists are used, which show the possibility of using a neural network to approximate arbitrary functions with a given accuracy. Thus, they showed a proof of the Hecht-Nielsen theorem, which defines the fundamental possibility of representing a continuous arbitrary function of many variables using a neural network with direct signal propagation when the network contains at least one layer of neurons. Structurally, such a neural network consists of N input neurons, at least from $2N + 1$ hidden neurons with sigmoidal activation functions of the form and M output neurons with unknown activation functions. It is proved that the parameters of the sigmoidal activation function can

Function name	Function Definition
Maximum subsampling	$\max[a \in A]$
Averaged subsample	$\frac{1}{ A } \sum_{a \in A} a$
Stochastic subsampling	$\sqrt{\sum_{a \in A} a^2}$
Probability	$p_i = \frac{a_i}{\sum_{a_j \in A} a_j}$

Table 1. Types of activation functions of a layers of a subsample of convolutional neural networks of A set

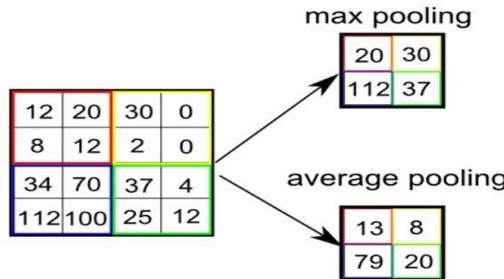


Fig. 1. Subsampling operation

be set a priori, and a linear activation function of the form can be used in the output layer of neurons:

$$g(x) = a + b$$

where g - linear activation function; x - the total input signal of the neuron, a, b - coefficients. Thus, neural network models of the multilayer perceptron type or neural networks with radial basis functions are subject to theoretical verification. Since the quality of recognition of aerial photography objects depends on the events that occur in it, and are characterized by a set of certain controlled parameters of computer vision, the model of detecting objects in aerial photographs can be written as follows:

$$\exists s(t) \in S_a(t) \wedge p(t) \in P_a(t) \Rightarrow A \quad (22)$$

$$\exists s(t) \in S_n(t) \wedge p(t) \in P_n(t) \Rightarrow A \quad (23)$$

$s(t)$ - a lot of pictures that appear in the information system; $p(t)$ - the set of values of the classification parameters of objects in the images in the information system at the

time t ; $S_n(t), P_n(t)$ - a lot of events in the information system and a lot of values of computer vision parameters characteristic of the implementation of signs of violation of the classification parameters of objects in aerial photographs; $S_a(t), P_a(t)$ - a set of events and a set of computer vision parameter values characteristic for determining the class of an object at a point in time t ; A - implementation of classification and identification of objects in aerial photographs.

Expressions (24, 25) can be supplemented by expressions (26, 27), which can be used to identify the normal state of object classification parameters on aerial photographs:

$$\exists s(t) \in S_a(t) \wedge p(t) \in P_a(t) \Rightarrow N \quad (24)$$

$$\exists s(t) \in S_n(t) \wedge p(t) \in P_n(t) \Rightarrow N \quad (25)$$

N - class normal state. The method of detecting objects on aerial imagery based on (2.23 and 2.24) is called "object class detection", and the method of object detection on aerial imagery based on (24 and 25) is called "anomaly detection" [5]. Using (26-27), a generalizing model for detecting objects in aerial photographs can be written as a continuous function of many variables:

$$\begin{cases} U = F(s(t), p(t), S_a(t), P_a(t), P_n(t)) \\ U \in (A, N) \end{cases} \quad (26)$$

As a result, the application of the Hecht-Nielsen theorem and the results of D. Park and I. Sandberg to the functional (28) allows us to state that with the help of a convolutional neural network, it is possible to recognize objects in aerial images in information systems with a given accuracy. At the same time, a necessary condition for the verification of the neural network model is the possibility of representing the parameters of object classification in the form of continuous functions.

One task of developing effective neural network tools for estimating the parameters of the classification of information system objects for identifying objects in aerial imagery is to determine the range of input parameters of the neural network model by convolution. The need to solve this problem is explained by the following factors:

- using as input parameters of the neural network model a large number of object classification parameters on aerial imagery images in the information system, which significantly increases the amount of computing resources and complicates the process of accumulating training examples;
- the use of uninformative parameters for classifying objects in aerial imagery leads to the training of a neural network model on noisy data, which negatively affects the correctness of the classification of unknown examples and increases the amount of computing resources;

- extraction of informative parameters of object classification in aerial imagery from the input parameters of the neural network model can lead to a complete loss of its classification properties.

At the same time, the final decision on the nomenclature of the input parameters of the neural network is made as a result of rather lengthy comparative experiments. To reduce the number of these experiments, it is advisable to determine the importance of each of the possible parameters of computer vision. Since modern formalized methods for assessing the importance of object classification parameters in aerial imagery do not meet the accuracy requirements, a decision was made to use expert evaluation to form the convolution parameters. It is proposed to use the method of paired comparisons, which is explained by its proven effectiveness in cases of a large number of test objects, with which the classification parameters of objects in aerial photographs are associated.

In this case, the input data of the model is a vector, the elements of which will be matrices of expert assessments of the significance of computer vision parameters.

As a result, a model has been developed for the processes of integrating object classification parameters in aerial imagery used to recognize unidentified objects and previously classified objects.

The developed model is used to determine the parameters of computer vision that can be used in neural network optimization tools for scanners for processing aerial imagery data based on the automation of expert evaluation of current parameters. A popular classifier choice is the Softmax classifier, which has a very different loss function and is based on a normalized exponential function. This classifier is essentially a generalized binary logistic regression classifier. Unlike the support vector machine, which computes the results of a vector $f(x_i, W)$ like estimates for each class, the Softmax classifier is driven by a slightly more intuitive approach and also has a probabilistic interpretation. In the Softmax classifier, the display loss function $f(x_i, W) = Wx_i$ remains unchanged, the cross-entropy function (29) is connected.

$$L_i = -\log \left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}} \right) \quad (27)$$

$$L(X, Y) = -\frac{1}{n} \sum_n^{i=1} \ln a(x^i) + (1 - y^i (\ln (1 - a(x^i))))$$

$X = \{x(1), \dots, x(n)\}$ — is a set of input examples for learning during input set. $Y = \{y(1), \dots, y(n)\}$ — this is the appropriate set of labels for input set. Function $a(x)$ — neural network output values with input value x .

According to the results, according to formula (30), an integral assessment of the state of the aerial survey area F of the entire territory is calculated:

$$F = \sum_i S_i \cdot LAI_i \cdot VQF_i \quad (28)$$

where adding a spectrum of object classification is carried out behind a class mask.

The mask is vectorized, within which the distribution of the integral indicator is vectorized. So, as a result, data for the cartographic presentation of the results has been prepared.

Findings. The aerial imagery time series analysis algorithm is based on the methods described in subsection 2.1.

The algorithm for analyzing the time series of multispectral satellite images consists of the following steps:

Step 1. Selecting images of the study area containing a time series.

Step 2. Preparation and pre-processing of selected images.

Step 3. Transformation of time series images to the spatial distribution of the indicator to be investigated - the number of object classes by LAI, or the quality of images by VQF, or a complex indicator.

Step 4. Processing the time series to obtain the parameters of the recognition trend for each class.

Step 5. Output of processing results.

The products obtained by the above algorithms of cartography are the basis for an integral quantitative and qualitative assessment of classes of objects of territories and the study of its changes over time. The final step is the development of current and future recommendations on the use of the developed methodology.

5. Conclusion

The article was a substantiated mathematical model and the main technologies of the described methodology and considered the general topology of the model of convolutional neural networks, also described all the components of their architecture and the values of parameters and hyperparameters of convolutional networks. The essence of the dropout technique was also described, which is widely used in preventing network retraining.

The article explains the essence of the normalization of input data for layers and the importance of preventing a decrease in the learning rate of the network and the number of parameters. Convolutional neural networks have found their application in image processing. The convolution operation alternates with the subsampling operation. Subsampling is applied to reduce the overall size image and increase the degree of invariance applied to it filters. When considering the network

architecture, we rely on the fact that the presence of some feature in the picture is much more important than the exact knowing its coordinates. Thus, the essence of the subsample is the choice the maximum neuron from several neighbors. Then this neuron is taken as an element of the next, but already reduced feature map. This operation helps to increase the scale invariance of the input Images. At the output of the convolutional layers of the network, several layers of a fully connected neural network are additionally installed that perform the work of the classifier.

REFERENCES

1. Buryachok V. L. et al. Features OF Implementation of the information security policy in the creation of the rational system of electronic documentary cooperation for public and commercial structures of ukraine //System technologies. – 2017. – №. 6. – C. 43-61.
2. Arifoglu D., Bouchachia A. Detection of abnormal behaviour for dementia sufferers using Convolutional Neural Networks //Artificial intelligence in medicine. – 2019. – T. 94. – C. 88-95.
3. Dokas P. et al. Data mining for network intrusion detection //Proc. NSF Workshop on Next Generation Data Mining. – 2002. – C. 21-30.
4. Khaikin S. Neural Networks: A Complete Course [Russian translation]. – 2008.
5. Abbaspour A. et al. Detection of fault data injection attack on uav using adaptive neural network //Procedia computer science. – 2016. – T. 95. – C. 193-200.
6. Zadeh L. A. Fuzzy sets as a basis for a theory of possibility //Fuzzy sets and systems. – 1978. – T. 1. – №. 1. – C. 3-28.
7. Weibull J. W. Evolutionary game theory. – MIT press, 1997.
8. Emel'yanova Y. G. et al. Neural network technology of detection network attacks on information resources //Program Systems: Theory and Applications. – 2011. – T. 2. – №. 3. – C. 3-15.
9. Komar M. et al. Compression of network traffic parameters for detecting cyber attacks based on deep learning //2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT). – IEEE, 2018. – C. 43-47.
10. Bernik I., Prislan K. Measuring information security performance with 10 by 10 model for holistic state evaluation //PloS one. – 2016. – T. 11. – №. 9. – C. e0163050.
11. Tchakoucht T. A. I. T., Ezziyyani M. Building a fast intrusion detection system for high-speed-networks: Probe and dos attacks detection //Procedia Computer Science. – 2018. – T. 127. – C. 521-530.

12. Ioffe S., Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift //International conference on machine learning. – PMLR, 2015. – C. 448-456.

UDC: 004.94

Multipath Redundant Transmission With Traffic Heterogeneity in Terms of the Criticality of Network Delays

V.A. Bogatyrev^{1,2}, A.V. Bogatyrev³, S.V. Bogatyrev^{2,3}

¹Department Information Systems Security Address, Saint-Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia

²ITMO University, Saint Petersburg, Russia

³JSC NEO Saint Petersburg Competence Center, Saint Petersburg, Russia

Abstract

The possibilities of increasing the probability of timely delivery of packets of heterogeneous traffic to the addressee are investigated as a result of setting the multiplicity of packet redundancy depending on their criticality to delays in the queues of switching nodes that make up the delivery path of a replica packet to the addressee.

For multipath redundant transmission of packets of heterogeneous traffic by acceptable packet delays in the network, an analytical model is proposed for estimating the probability of timely delivery of at least one of the replicas of packets transmitted to the addressee. The efficiency of destroying expired packets in intermediate nodes that make up data transmission paths is shown.

For heterogeneous traffic, the existence of an optimal multiplicity of redundancy of packets of various streams is shown, determined depending on the criticality of the network delays allowed for them.

Keywords: heterogeneous traffic, redundancy, multipath transmissions, real-time, packet replication

1. Introduction

The functioning of distributed computer systems in real time requires their high reliability and fault tolerance [1-4] with low delays for data transmission and processing. In connection with the intensive development of cyberphysical systems [5], increased attention is paid to the solution of these tasks, which was also reflected in the creation of the concept of Ultrareliable and Low-Latency Wireless Communication [6-7].

High reliability of transmissions in the network is achieved with the introduction of continuity, including on the basis of noise-resistant coding and establishing

connections with the issuance of confirmations and the implementation of repeated transmissions. (for example, the TCP transport protocol), which is associated with an increase in channel loading and packet transmission delays. The Real-Time Transport Protocol (RTP), used together with the Resource Reservation Protocol (RSVP), is aimed at ensuring the timeliness of transfers without confirmation of delivery and retransmissions.

The multipath implementation of the TCP Multipath TCP (MPTCP) protocol allows you to reduce delays and increase transmission throughput. For a single Multipath TCP connection, interacting nodes can use different IP addresses to exchange packets [8, 9]. The timeliness of MPTCP transmissions can be reduced if there are tight policy deadlines for acceptable network delays, which are difficult to meet when issuing acknowledgments.

The timeliness of packet delivery without acknowledgments and retransmissions is ensured by the joint use of the Real-Time Transport Protocol (RTP) and the Resource Reservation Protocol (RSVP) [10,11]. Transport coding [12, 13] is aimed at reducing average delays in the network, assuming fragmentation of messages with coding that allows restoring their content in case of distortion or non-delivery of some fragments. Transport coding, while minimizing retransmissions, at the same time complicates the processing of packets on the receiving and transmitting sides, in connection with encoding, decoding and message recovery.

The development of the idea of encoding with preventive error correction (Forward Error Correction, FEC) is implemented in the QUIC protocol (Quick UDP Internet Connections). The QUIC protocol [18,19] allows multiplexing of several data streams between two computers, working on top of the UDP protocol. In the QUIC protocol, each transmitted packet is supplemented with data from other packets, which makes it possible to recover lost and corrupted packets from the contents of the completed transmissions. Information recovery is implemented without retransmissions of the lost packet, however, in this case, the packet redundancy increases by about 10%. The implementation of the QUIC protocol, as well as transport coding, is associated with the complexity of calculations on the transmitting and receiving sides.

If real-time operation limits the possibility of retransmissions within the maximum allowable time, then (in addition to the considered approaches with the introduction of information redundancy), especially with a high level of errors in the channels, the reliability and timeliness of transmissions in structurally redundant infocommunication systems can be increased as a result of multipath redundant transmissions

The multipath redundant transmission provides for packet replication with a task for each replica of the path of its delivery to the addressee [16-18]. This requires timely delivery to the addressee of at least one of the generated replicas. The transmitted

packets reflect the allowable time of their stay in the network. In intermediate nodes, the waiting time for replicas in their queue is calculated, and the value of the field that reflects the allowable waiting time is reduced by this value. Packets for which the accumulated total waiting time in the passed nodes queues exceeds the time limit are discarded. The destruction of expired (irrelevant) replicas in intermediate nodes, taking into account the accumulation of delays along the way, complicates the assessment of the probability of timely delivery to the addressee of at least one formed replica of the transmitted packet [20].

The organization of redundant multipath transmissions is a development of the concept of multipath routing [19,20], but unlike it, it involves the formation of replicas of packets with their transmission along different routes (paths). The effective use of redundant transmissions requires the resolution of a technical contradiction associated with a possible increase in the probability of on-time delivery of at least one replica and, at the same time, its decrease due to an increase in the overall load due to replication. Thus, it is required to determine the area of expedient use of redundant multipath transmissions depending on the traffic intensity, the frequency of packet replication and their allowable delays in the network.

Models for estimating the probability of timely, error-free delivery of packets to the addressee with multipath transmission reservation during the passage of replicas of the transmitted packets of the sequence of switching nodes that make up the path for traffic that is homogeneous in terms of permissible delays are proposed in [16, 18].

Modern infocommunication systems are characterized by the heterogeneity of the flow of requests in terms of functionality and criticality to reliability and acceptable transmission delays in the network. Such heterogeneity of traffic makes it necessary to control the provision of network resources to transmitted packets, depending on the criticality of delays in their delivery to the addressee.

The purpose of the work is to research the possibilities of increasing the probability of timely delivery of heterogeneous traffic packets as a result of multi-hop transmissions with the multiplicity of packet replication, depending on their criticality to network delays when destroying irrelevant (expired) packets in intermediate communication nodes.

When the flow is heterogeneous in terms of admissible delays in the network, the determination of the optimal multiplicity of their replication is largely due to the choice of the integral indicator of the efficiency of heterogeneous traffic transmissions. The rationale for the integral indicators for traffic heterogeneity in terms of admissible delays is given in [17], however, it does not study the influence of the destruction of irrelevant replicas in intermediate nodes on the choice of the redundancy ratio of transmissions critical to delays.

This article is devoted to the analysis of the impact on the probability of timely delivery of heterogeneous traffic by setting the replication rate of transmitted packets depending on their allowable delays in the network, taking into account the destruction of expired replicas in intermediate nodes.

2. Probability of timely delivery of heterogeneous traffic packets with transmission reservation

In case of redundant transfers of heterogeneous traffic according to the allowable packet delays in the network, the packet replication rate can be set depending on the delays t_0 allowed for them.

If there are n possible non-intersecting data transfer paths in the network, then the intensity of requests sent along each path will be $\Lambda = \Lambda_S/n$, where Λ is the total intensity of requests entering the system.

Let traffic inhomogeneities be distinguished by z gradations of criticality of requests to acceptable waiting in the network. In the case of transmissions with a packet replication rate, depending on their criticality to network delays, for switching nodes that receive the packets formed replicas first (located at the beginning of the path, $i = 1$), the intensity of the incoming flow is

$$\Lambda_1 = \Lambda \sum_{b=1}^z \beta_b k_b,$$

where β_b is the share of the b -th stream, k_b is the multiplicity of reservation of transmissions of the b -th stream.

$$\sum_{b=1}^z \beta_b = 1.$$

When the expired packets are destroyed in the intermediate nodes of the path, the intensity of requests arriving at the input of the i -th node of the path ($i > 1$) for heterogeneous traffic is defined as

$$\Lambda_i = \Lambda \sum_{b=1}^z k_b \beta_b \prod_{j=1}^{i-1} P_{jb},$$

In this case, P_{ib} is the probability of not exceeding the allowable waiting time for a packet of the b -th flow when passing through the i -th node, taking into account the average delays in the queues of previously passed switching nodes. We represent communication nodes as single-channel queuing systems with an infinite queue of the $M/M/1$ type [21, 22], then for $i > 1$

$$P_{ib} = (1 - \Lambda_i v_i e^{(\Lambda_i - \frac{1}{v_i})(t_b - \sum_{j=1}^{i-1} w_j)}),$$

and for $i = 1$

$$P_{1b} = (1 - \Lambda_1 v_1 e^{(\Lambda_1 - \frac{1}{v_1})t_b}),$$

at the same time, for the j -th switching node, w_j is the average waiting time for a packet (its replica) in the queue, and v_j is the average request service time

$$w_j = \frac{\Lambda_j v_j^2}{1 - \Lambda_j v_j}.$$

The probability P_b of timely delivery to the addressee of a certain packet of the b -th flow when passing m nodes that make up the path is defined as

$$P_b = \prod_{i=1}^m P_{ib}.$$

When reserving transmissions of heterogeneous traffic, the desired probability R_b of timely delivery to the addressee of at least one replica of the packet of the b -th flow (at least through one of the k_b paths involved for this) is calculated as

$$R_b = 1 - (1 - P_b)^{k_b}.$$

For heterogeneous traffic, the mathematical expectation of the probability of timely delivery to the addressee of packets of different criticality to the waiting delays at the nodes of the path is defined as

$$P_c = \sum_{b=1}^z \beta_b R_b.$$

However, this indicator does not take into account the importance of timely delivery of packets of various criticality to network delays and does not allow assessing income and risks that depend on the fulfillment of the conditions for their timely delivery to the addressee.

The generalized indicator of the efficiency of transfers of traffic that is heterogeneous in terms of permissible delays should be determined based on the characteristics of the applied problems solved by the system and take into account the risks associated with the untimely delivery of packets of different streams [17].

The efficiency of the system functioning can be characterized by the profit from the provision of timely transmission services in the network [17]. Let the profit from

the timely delivery of the package of the j -th flow be c_j , and the penalties for late delivery s_j , then the mathematical expectation of the profit from the execution of the request of the total flow will be [17]

$$C = \sum_{j=1}^z \beta_j (R_j c_j - (1 - R_j) s_j).$$

The mathematical expectation of profit per unit of time (intensity of profit) can be defined as [17]

$$D = \Lambda \sum_{j=1}^z \beta_j (R_j c_j - (1 - R_j) s_j).$$

3. Results of calculating the probability of timely delivery of packages

When calculating, will assume that each path passes through $m = 3$ switching nodes, the average transmission times through which (excluding waiting in queues) are $v_1 = v_2 = v_3 = 0.1$ s. With traffic heterogeneity, we single out two gradations of packet criticality to delays in the sequence of switching nodes that make up the path of their delivery to the addressee.

In Fig.1. the dependence of the mathematical expectation of the probability of timely delivery of packets on the intensity of heterogeneous traffic with the same multiplicity of reservation of all packets is presented. Curves 1–3 correspond to the multiplicity of transmission reservation for all requests of a heterogeneous flow $k = 1, 2, 3$ when the expired replicas of packets in intermediate nodes are destroyed, and curves 4–6 - without their destruction at $k = 1, 2, 3$. The calculations are performed for $\beta = 0.8$, $v_1 = v_2 = v_3 = 0.1$ s and $t_1 = 0.4$ s, $t_2 = 0.8$ s.

In Fig.2. the dependences of the average probability of timely delivery of packets to the addressee on the intensity of heterogeneous traffic are presented when replicating only delay-critical packets. Curves 1, 2, 3 correspond to the transmission of delay-critical packets during their replication with multiplicity $k = 1, 2, 3$, when expired packets are destroyed in intermediate switching nodes. Curves 4, 5, 6 correspond to the transmission of delay-critical packets with their replication multiplicity $k = 1, 2, 3$ without destroying expired requests. Calculations were performed at $\beta = 0.5$, $v_1 = v_2 = v_3 = 0.1$ s and $t_1 = 0.2$ s, $t_2 = 1$ s.

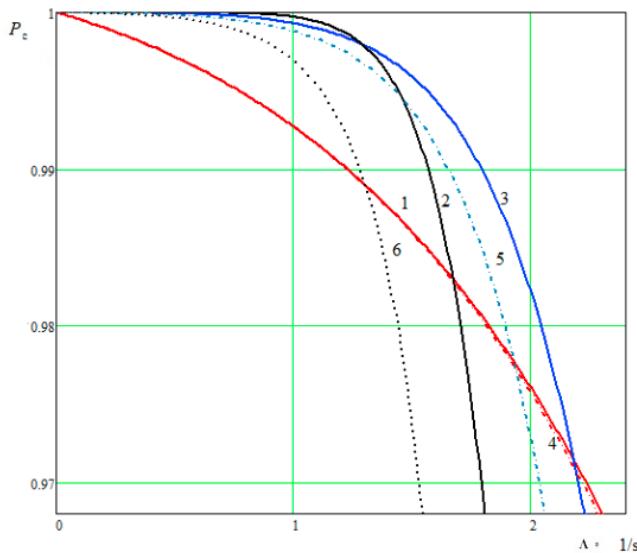


Fig. 1. The dependence of the mathematical expectation of the probability of timely transfers on the intensity of heterogeneous traffic with the same multiplicity of reservations for all packets.

The calculations have been performed in the computer mathematics system «Mathcad 15».

The presented graphs confirm the high efficiency of the destruction of expired packets in intermediate nodes and the existence of an optimal multiplicity of reservation (replication) of packets that are critical to the total delay in the queues of the nodes transmitting them. As the system load decreases, the redundancy ratio, at which the maximum average probability of timely delivery of at least one of the replicas of the transmitted packets is reached, increases.

The figures show the expediency of an adaptive change in the multiplicity of packet replication with an increase in traffic intensity. So, at low traffic intensity, it is necessary to reserve transmitted packets with a multiplicity of three, as it increases, with a multiplicity of two, while there is a traffic intensity limit, above which it is not advisable to reserve packets.

The figures show the existence of an optimal multiplicity of reservations that are critical to the total delay in the queues of requests. Moreover, when the requests expired in the queues of the switching nodes, the effect of transmission reservation increases.

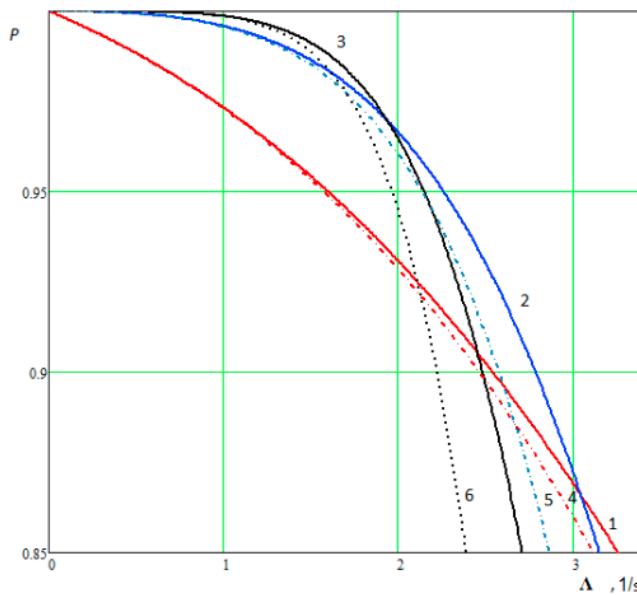


Fig. 2. The dependence of the mathematical expectation of the probability of timely transfers on the intensity of non-uniform traffic when reserving only delay-critical packets.

The presented dependencies show the expediency of reserving transmissions of the most delay-critical packets of a heterogeneous stream without replicating the packets that are less delay-critical.

4. The results of calculating the profit from the timeliness of the transmission of heterogeneous traffic

Let us analyze the dependence of the profit D , received per unit of time, during the transmission of heterogeneous traffic on its total intensity Λ .

Consider the case when, during transmission, only packets with a lower allowable waiting time in the queues of the nodes that make up the path are reserved. We assume that the profit from the timely delivery of the package of the first and second streams $c_1 = 10$ c.u., $c_2 = 5$ c.u., and the penalties for late delivery $s_1 = s_2 = 40$ c.u.

Calculations were performed at $v_1 = v_2 = v_3 = 0.2$ s, $t_1 = 1$ s, $t_2 = 2$ s and $\beta = 0.7$. The dependence of profit on the intensity of the flow Λ is shown in Fig. 3, in which the multiplicity of reservations of packets critical to delays in the network $k = 1, 2, 3$ correspond to curves 1-3 when destroying overdue packets along the way, and without destroying them - curves 4-6.

When destroying packages overdue in transit, Fig. 4 shows the dependence of profit on the intensity of the flow Λ . Curves 1 - 3 correspond to the multiplicity of

redundancy critical to the delivery of packages $k = 1, 2, 3$ at $\beta = 0.5$, and curves 4 - 6 at $\beta = 0.2$.

The figures show the effectiveness of the destruction of overdue packets along the way and the feasibility of increasing the multiplicity of reservations of packets that are critical to delays in the network as the traffic intensity decreases. For heterogeneous traffic, it can be concluded that there is an optimal multiplicity of packet reservations, depending on the criticality of the network delays allowed for them. The proposed models and technical solutions for ensuring the reliability and timeliness of multipath redundant transmissions of heterogeneous traffic in terms of allowable delays in the queues of network nodes are supposed to be adapted for use within wireless networks within the concept of Ultrareliable and Low-Latency Wireless Communication [23, 24], including systems that do not allow interruptions of the computational process during recovery.

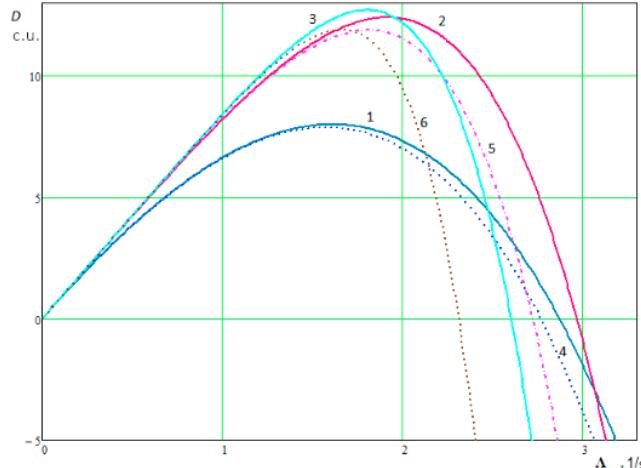


Fig. 3. Profit D from the transmission of heterogeneous traffic when reserving packets that are critical to network delays with and without destruction of overdue packets along the path.

5. Conclusion

An analytical model is proposed for estimating the probability of timely redundant transmissions, heterogeneous traffic in networks that involve replication of transmitted packets with a multiplicity depending on their criticality to the allowable waiting time in the network. The model takes into account the possibility of destruction in intermediate nodes of the path of replicas of transmitted packets, the accumulated waiting time of which in the path exceeded the established limit.

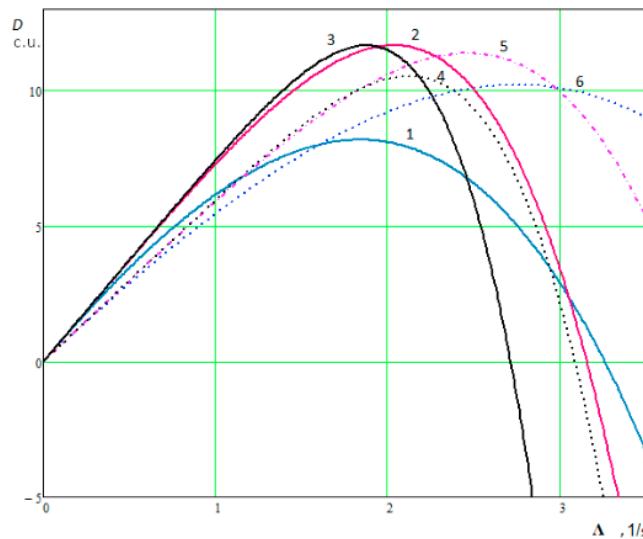


Fig. 4. Profit D from the transmission of heterogeneous traffic when reserving packets that are critical to network delays at $\beta = 0.5$ and $\beta = 0.2$.

It is shown that the reservation of the most critical packets in terms of the allowable waiting time allows increasing the average probability of timely delivery of the entire set of packets and increasing the intensity of the profit received in the network in total from the transmission of heterogeneous traffic. It is shown that when expired packets are destroyed in transit, the effect increases.

For heterogeneous traffic, the existence of an optimal multiplicity of packet reservations is shown, which is determined depending on the criticality of network delays allowed for them and on the total intensity of heterogeneous traffic.

REFERENCES

1. Aysan H., Fault-tolerance strategies and probabilistic guarantees for real-time systems Mälardalen University, Västerås, Sweden. 2012. 190 p.
2. Kim S., Choi Y., Constraint-aware VM placement in heterogeneous computing clusters. Cluster Comput 23, 71–85 (2020). <https://doi.org/10.1007/s10586-019-02966-6>.
3. Tatarnikova T.M., Poyanova E.D., Differentiated capacity extension method for system of data storage with multilevel structure. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2020, 20(1), . 66–73 DOI: 10.17586/2226-1494-2020-20-1-66-73

4. Bennis M., Debbah M., Poor H.V., Ultrareliable and Low-Latency Wireless Communication: Tail, Risk and Scale, Proc. IEEE 2018, 106, pp.1834–1853. doi: 10.1109/JPROC.2018.2867029.
5. Ji H.; Park S., Yeo J., Kim Y., Lee J., Shim B., Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects, IEEE Wirel. Commun., 2018, 25, pp.124–130. doi :10.1109/MWC.2018.1700294.
6. Sachs J., Wikström G., Dudda T., Baldemair R., Kittichokechai K., 5G Radio Network Design for Ultra-Reliable Low-Latency, Communication IEEE Netw., 2018, 32, pp. 24–31. doi:10.1109/MNET.2018.1700232.
7. Pokhrel S., Panda M., Vu H., (2017-02-24). "Analytical Modeling of Multipath TCP Over Last-Mile Wireless". IEEE/ACM Transactions on Networking. 25 (3): 1876–1891. doi:10.1109/TNET.2017.266352
8. Peng Q., Walid A., Hwang J., and Low S. H. 2016. Multipath TCP: Analysis, design, and implementation. IEEE/ACM Transactions on Networking 24, 1 (2016),596–609.
9. Perkins C., RTP— Addison-Wesley, 2003. — P. 414. — ISBN 9780672322495.
10. Zurawski R., RTP, RTCP and RTSP protocols // The industrial information technology handbook CRC Press, 2004. — P. 28—70. — ISBN 9780849319
11. Krouk E., Semenov S., Application of Coding at the Network Transport Level to Decrease the Message Delay // Proc. of 3rd Intern. Symp. on Communication Systems Networks and Digital Signal Processing. Staffordshire University, UK, 2002. P. 109—112
12. Kabatiansky G., Krouk E., Semenov S., Error Correcting Coding and Security for Data Networks. Analysis of the Superchannel Conc ept. Wiley, 2005. 288
13. De Coninck Q.; Bonaventure O. (2010-12-12). "Multipath QUIC: Design and Evaluation". Proc. Conext'2017, Seoul, Korea.
14. Roskind J.. 2013. QUIC(Quick UDP Internet Connections): Multiplexed Stream Transport Over UDP. Technical report, Google 2013
15. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V., Redundant Servicing of a Flow of Heterogeneous Requests Critical to the Total Waiting Time During the Multi-path Passage of a Sequence of Info-Communication Nodes. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020. Vol. 12563. pp. 100-112
16. V.A. Bogatyrev, A.V. Bogatyrev, S.V. Bogatyrev, The probability of timeliness of a fully connected exchange in a redundant real-time communication system, Wave Electronics and its Application in Information and Telecommunication Systems (WECONF 2020). <https://ieeexplore.ieee.org/document/9131517>. doi:10.1109/WECONF48837.2020.9131517

17. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V., Multipath redundant real-time transmission with the destruction of expired packets in intermediate nodes CEUR Workshop Proceedings, 2021, 3027, pp. 971–979
18. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V., Inter-machine exchange of real time in distributed computer systems CEUR Workshop Proceedings, 2020, 2744
19. Merindol P., Improving Load Balancing with Multipath Routing / Merindol P., Pansiot J., Cateloin S. // Proc. of the 17-th International Conference on Computer Communications and Networks, IEEE ICCCN 2008. – 2008. – P. 54-61.
20. Prasenjit C. , Tuhina S., Indrajit B., Fault-tolerant multipath routing scheme for energy efficient wireless sensor networksInternational Journal of Wireless Mobile Networks (IJWMN) Vol. 5, No.2,April 2013 pp 33-45
21. Kleinrock L., Queueing Systems: Volume I. Theory. New York: Wiley Interscience.1975 p. 417. ISBN 978-0471491101.
22. Kleinrock L., Queueing Systems: Volume II. Computer Applications. New York:Wiley Interscience. 1976 p. 576. ISBN 978-0471491118.
23. Ji H.; Park S., Yeo J., Kim Y., Lee J., Shim B., Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects. IEEE Wirel. Com-mun. 2018, 25, 124–130.
24. Sachs J., Wikström G., Dudda T., Baldemair R., Kittichokechai K., 5G Radio Network Design for Ultra-Reliable Low-Latency Communication. IEEE Netw. 2018, 32, 24–31.

UDC: 004.82; 621.391

IIoT information processing model for transfer learning with data quality management

A.Y. Grebeshkov¹

¹Povelzhskiy State University of Telecommunications and Informatics, 443010,
L.Tolstoy str., 23, Samara, Russia

grebeshkov-ay@psuti.ru

Abstract

Generally, the use of transfer learning techniques is challenging when the testing data and the training data are not collected into a single enterprise data storage and the data items are not linked explicitly one to another, as happens with the sensor data flows in the sensor network. The objective of this paper is to propose a transfer learning service delivery model based on the collaborative information processing with consolidation data sets and data quality management. In order to achieve objective we suppose model with an ontology-based rules applying to check quality of an enterprise database which is used in Industrial internet of things platform for machine learning service support.

Keywords: Industrial internet of things, data quality, machine learning, transfer learning, sensor network, IIoT, ML

1. Introduction

Traditional machine learning techniques try to provide machine learning task or service from a specific data set. This data set can be considered as a big data, but data set has restrictions on the data source quality. On the contrary, a transfer learning techniques try to transfer the knowledge from some previous machine learning results to a target task or service. This approach take place when the target task has fewer high-quality training data [1]. The industrial application of the transfer learning allows that training data and testing data not to be collected into a single virtual storage. With providing reproducible results of machine learning we implement the transfer learning on the base of collaborative information processing in intelligent sensor networks.

2. Transfer learning and data-quality management

Transfer learning techniques using is ideal where training data is expensive or time-consuming. Since for a machine learning based service is a reason to use abundant labeled training data sets, which have the same distribution as testing data set. There are about 40 transfer learning approaches that are divided into homogeneous and heterogeneous transfer learning [2; 3]. Homogeneous transfer learning is for handling the scenario where the subject-oriented or problem-oriented domains are of the same feature space. Heterogeneous transfer learning process take into account the knowledge transfer process between domains with different feature spaces. It is need to understand that an information object may have different meaning context in the case of homogeneous or heterogeneous transfer learning. This point is a key in a case of reinforcement learning tasks that characterized by sequential decision making and delayed reward and where successfully transfer knowledge represented as an expected reward from each state [4].

The heterogeneous transfer learning scenario supposes that feature spaces between the source and the target are nonequivalent and there is a handling the cross-domain differences with symmetric or asymmetric feature transformation including semantic knowledge transfer without detailed context analysis [5]. Transfer learning process and knowledge transfer are discussed as a part of data-driven memetic computation [6], and for treating image and text to improve text processing techniques with semantic linking between text and image [7]. There is a method combines transfer learning and semi-supervised learning when training and testing data has different distributions [8]. There is an interactive environment that supports transfer learning with selection, assembly and diagnostic model challenges [9]. Description of sequential learning assuming the availability of labeled data as in the source domain as in the target domain, where the source domain is handled first can be found in [10]. There are two base scenarios with centralized transfer learning and distributed transfer learning for IIoT systems with different data sets without cross-domain differences in [11].

It is newer appropriate to apply the ontology-based data quality control technique is described in the context of data management for handling the cross-domain differences or for the train data set and the test data set processing with different distributions [12; 13]. In general this approach helps to validate information objects extracted from various data sets and to represent data objects according to different ontologies reflecting various domain. This approach allows to make identification of the object classes provided by domains. The result is a set of data objects for each domain, variables and their relations.

The first stage of the proposed method is a subject-oriented ontology composing with a labeled cross-domain entity relations. The next stage is mapping rules using

to define relations between elements of ontology and data set or database parameters like as conceptual data model with variables. Data set is a training data storage, database is a testing data storage for ML-based IIoT service providing. Individuals of data set and database classes may matching and these individuals would be base for data set and database consolidation. For further research it is need to apply special program database adapter at the data access layer to extract individuals matching each variables from database.

As result there is a set of data objects and their relations and variables. Next procedure is a transformation into unified form to handle by data abstraction layer. Data-quality management model using a data abstraction layer to represent data schema and mapping rules. It can be built over existing industrial internet of things platforms, data sets and databases without modifying them. The data abstraction layer should provide an application program interface to query data from databases or logical data mart, to perform learning tasks, and to consolidate data objects between data sets and databases.

The data model is being developed where data objects from various databases and data sets have the same context, same key or label, and can be consolidated in the virtual data set as a single data objects set for future ML based IIoT service providing at the upper layer data-quality management. The continuing use of the consolidate data for machine learning includes standard procedures as a format consistency of data records, reduce data, data cleaning, re-scale data and others procedures beyond areas of this paper. Finally, cleared and reduced database can be considered as a data set for transfer learning.

3. IIoT information processing model for transfer learning

Proposed IIoT information processing model is used to represent the general components for transfer learning, to request checking data quality and to consolidate data, wherever data stored (see Fig. 1). This model supports the collaborative information processing in the intelligent sensor networks based on the ISO/IEC 20005:2013 standard.

The R denotes relation between the m component of the model and the n component of the model. This relation can be set out as equation (1):

$$mRn \rightleftharpoons \langle m, n \rangle \in R. \quad (1)$$

The m, n pair can be regarded as an ordered pair, where the first element in the pair is assigned the role of ML-based IIoT service user, the second element in the pair is assigned the role of ML-based IIoT service provider or data manager provider and this pair's relation can be estimated with equation (2) :

$$\langle m, n \rangle \rightleftharpoons \{\{m\}, \{m, n\}\}. \quad (2)$$

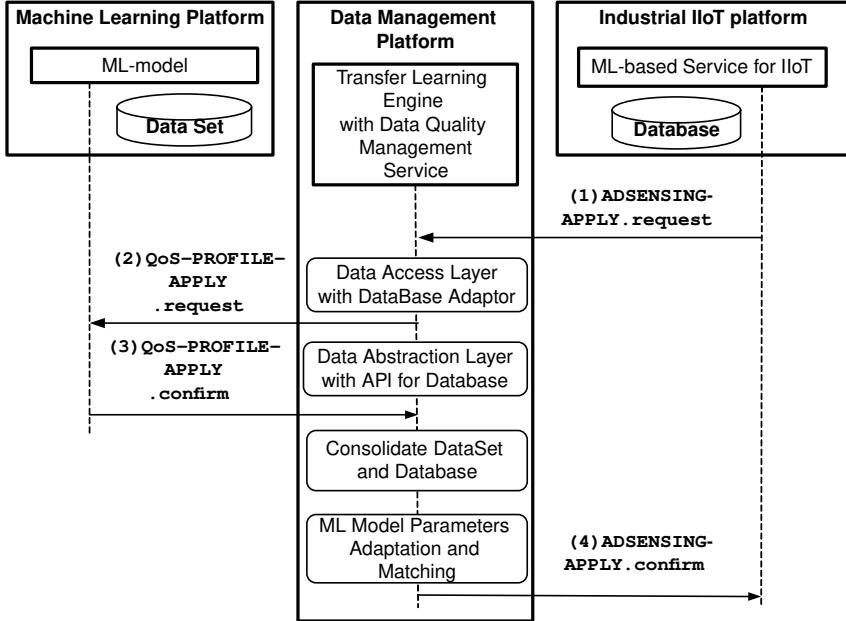


Fig. 1. IIoT information processing model for transfer learning with data quality management

Let's the SML predicate denotes a property of the m, n pair which is described as a "ML-based service relationship". If the R set is exist as follows:

$$R = \{\langle m, n \rangle \mid SML(m, n)\}, \quad (3)$$

it is also true that:

$$SML(m, n) \Leftrightarrow mRn. \quad (4)$$

There is the ML set of modeling components and there is at least one m, n pair belongs to the ML set for which there is the $\Gamma_m^{req}(ML)$ set of requests is described as follows:

$$\Gamma_m^{req}(ML) := \{mRn \mid n \in ML \setminus \{m\}\}, \quad (5)$$

where the req denotes request for ML-based IIoT service providing.

In Fig 1. the *ADSENSING* adaptive sensing service request is provided through service access point as a logical interface which incorporates the *ADSENSING-APPLY.request* primitive and the *ADSENSING-APPLY.confirm* primitive, which can be written in the form (6). This ML-based IIoT service request supports a ML-features of industrial IIoT platform linked with sensor system's perception for a physical world or for parameters of an industrial equipment.

$$\begin{aligned} \text{ADSENSING - APPLY.request} \{ & \\ & \text{ADSENSINGRequestorID}, \\ & \text{ADSENSINGTargetID}, \\ & \text{ADSENSINGMLList}, \\ \} & . \end{aligned} \quad (6)$$

The *ADSENSINGRequestorID* parameter is an identity of ML-based IIoT service user, the *ADSENSINGTargetID* parameter is an identity of data-management platform as a ML-based IIoT service provider. It is proposed to add the *ADSENSINGMLList* parameter with ML-based service descriptions and conceptual data model descriptions. There is the *ADSENSINGMLResultCode* parameter in *ADSENSING-APPLY.confirm* with results of checking rules applying and ML-model parameters.

The *QOS-PROFILE* QoS sensing service is provided through service access point as a logical interface which incorporates the *QOS-PROFILE-APPLY.request* primitive and the *QOS-PROFILE-APPLY.confirm* primitive. This service supports a features concerning update or changes QOS IIoT service due to machine-learning technique applying. The ML-based IIoT service can be reproduced as repository or machine learning model mart with Machine Learning Platform support. Primitive can be written in the form (7).

$$\begin{aligned} \text{QoS - PROFILE - APPLY.request} \{ & \\ & \text{QoSRequestorID}, \\ & \text{QoSProfileManagerID}, \\ & \text{QoSProfileMLList} \\ \} & . \end{aligned} \quad (7)$$

The *QoSRequestorID* parameter is an identity of user for ML-based IIoT service, the *QoSProfileManagerID* parameter is a description of stages sequences which was used in Section 2. It is proposed to add the *QoSProfileMLList* parameter that includes mapping rules. There is a modification the the *QoSResultCode* parameter in *QoS-PROFILE-APPLY.confirm* primitive with transfer learning model parameters and labeled data set object using with ML-service database objects.

4. Conclusion

Ontology-based data quality management is convenient for delivering of transfer learning techniques with data quality control according to the subject area's conceptual model, regardless of the physical data and sensor network structures. The proposed information processing model may supplement the Data Management Platform features with support of collaborative information processing for ML-based IIoT service with transfer learning.

REFERENCES

1. Sinno J. P., Qiang Y. A survey on transfer learning //The Knowledge Engineering Review. 2010. V. 22. No. 10. P. 1345–1359.
2. Zhuang F., Qi Z., Duan K. et al. A comprehensive survey on transfer learning //Proceedings of the IEEE. 2021. V. 109. Iss. 1. P. 43–76.
3. Yang Q., Zhang Y., Dai W., Pan S. J. Transfer Learning. Cambridge university press, 2020.
4. Taylor M. E. Transfer in reinforcement learning domains. Springer-Verlag, 2009.
5. Day O., Khoshgoftaar T. M. A survey on heterogeneous transfer learning //Journal of Big Data. 2017. Vol. 4. No. 29. P. 1–42.
6. Gupta A., Ong Y.-S. Memetic computation. Springer Nature, 2019.
7. Ionescu R.T., Popescu M. Knowledge transfer between computer vision and text mining. Springer international publishing, 2016.
8. Zhou H., Zhang Y., Huang D., Li L. Semi-supervised learning with transfer learning. In: Sun, M., Zhang, M., Lin, D., Wang, H. (eds) Chinese computational linguistics and natural language processing based on naturally annotated big data. LNCS 8202. Springer. 2013. P. 109–119.
9. Mishra S., Rzeszotarski J. M. Designing interactive transfer learning tools for ML non-experts // Proc. of the 2021 CHI Conference on human factors in computing systems (CHI'21). 2021. P. 1–15.
10. Wong L. J., Michaels A. J. Transfer learning for radio frequency machine learning: a taxonomy and survey //Sensors. 2022. Vol. 22. 1416. P. 1–14.
11. Liu X., Yu W., Liang F. et al. Towards deep transfer learning in industrial internet of things // IEEE Internet of things journal. 2021. Vol. 8. Iss. 15. P. 12163–12175.
12. Daraio C., Lenzerini M., Leporelli C. et al. The advantages of an ontology-based data management approach: openness, interoperability and data quality // Scientometrics. 2016. Vol.108. Iss. 1. P. 441–455.
13. Gorshkov S., Grebeshkov A., Shebalov R. Ontology-based industrial data management platform, <https://arxiv.org/abs/2103.05538>

УДК: 004.75

Теория и практика определения уровня критичности инцидентов в цифровых инфраструктурах

К.И. Михайлов, А.Г. Абрамов¹

¹Санкт-Петербургский политехнический университет Петра Великого,
Политехническая ул., д. 29, Санкт-Петербург, Россия

k.mikhaylov@gmail.com, abramov_ag@spbstu.ru

Аннотация

В статье рассматриваются некоторые прикладные аспекты мониторинга и управления цифровых инфраструктур с акцентом на трудностях, возникающих при определении уровня критичности событий (инцидентов). Выделяются основные источники и механизмы поступления от компонентов инфраструктуры в системы мониторинга сообщений (первичных событий), методы их классификации и формирования вторичных событий (инцидентов), а также оценки уровня критичности. Кратко представляется архитектура программной платформы мониторинга и управления «Центральный Пульт», обсуждаются принципы работы с событиями и инцидентами, доступные методы определения уровня критичности.

Ключевые слова: цифровая инфраструктура, мониторинг и управление, программный комплекс, иерархия объектов, связанность, событие, инцидент, уровень критичности

1. Введение

В условиях постоянно возрастающих в последние декады роли и значения информационно-коммуникационных технологий (ИКТ) в бизнес-процессах организаций из широкого спектра отраслей, увеличивающейся зависимости функционирования производств и оказания услуг от непрерывности и надежности работы средств автоматизации деятельности, сложных цифровых экосистем на первый план выходят технологии и программные решения мониторинга и оперативного управления (см., например, [1, 2, 3] и ссылки в них).

Системы мониторинга и управления нацелены на выполнение в постоянном режиме комплексного «аудита» цифровых инфраструктур (включая отдельные элементы, комплексы и услуги), отвечают за процессы управления ИКТ-средами

и консолидацию информационных потоков, формируя единую среду для обмена, сквозной интеграции и оркестрации данных, реализуя механизмы сбора информации из разных источников, обработки и интеллектуального анализа [4].

Одна из значимых составляющих процесса мониторинга связана с эффективной организацией управления происходящими в инфраструктуре инцидентами, которые представляют собой регистрируемые и классифицируемые системой мониторинга события, требующие повышенного внимания и своевременной реакции [5, 6]. Система мониторинга и управления в идеале должна однозначно и правильно классифицировать событие, присвоить (или подтвердить) уровень критичности (важности, опасности, значимости - severity), обоснованно принять решение о приоритетности и способе реагирования и, при наличии возможности, реализовать его, в том числе с применением средств автоматизации (автоматически устранить причину сбоя или предпринять действия, предупреждающие сбой в предоставлении сервисов или деградации качества). Полная автоматизация процессов и операций сборки, развертывания, масштабирования, мониторинга и управления рассматривается в профильном сообществе как концепция роботизированных операций (Artificial Intelligence for IT Operations, AIOps).

Большое внимание традиционно уделяется разработке и совершенствованию методик и инструментов мониторинга и управления инцидентами информационной безопасности (SIEM/IRP/SOAR/SGRC). Отдельные прикладные вопросы, относящиеся к определению уровня критичности инцидентов в ИКТ-средах, отражены в документации, доступной на сайтах ведущих разработчиков систем мониторинга и управления (в том числе таких как DX Spectrum, IBM Tivoli, SolarWinds, PRTG Network Monitor, Microsoft SCOM, Zabbix).

2. Источники данных мониторинга, события и инциденты

Цифровая инфраструктура в процессе своего функционирования генерирует поток сообщений (первичных событий), которые могут поступать в систему мониторинга на основе различных методов, транспортов и сетевых протоколов.

Системы мониторинга позволяют получать сообщения, направленные из файлов журналов сетевого и серверного оборудования, виртуализованных сред, прикладных сервисов (syslog), SNMP-ловушки («traps»), сетевые потоки NetFlow, зеркальную копию IP-трафика и др. Отдельный интерес представляют специальные протоколы/форматы данных, среди которых - оригинальные, кастомные (например, оборудования для мобильной связи), MQTT-события от устройств Интернета вещей (Internet of Things, IoT) и контроллеров автоматизации, события на основе открытых протоколов машинного взаимодействия Modbus, ProfiBus. Нельзя не упомянуть здесь о таких источниках событий, как сообщения и метрические данные, поступающие от приложений - от кодовых

включений, указанных авторами в тексте программ, и косвенным образом собранных данных о том, насколько быстро и качественно программы работают (класс решений APM, Application Performance Management).

Еще одним важным источником событий от инфраструктуры являются данные, поступающие от локальных или специализированных систем управления. Например, телекоммуникационные комплексы, комплексы управления сетями хранения данных, распределенными вычислениями самостоятельно выполняют консолидацию и классификацию событий, выдают поток данных о происходящих в локально сконцентрированной части инфраструктуры событиях и инцидентах.

Во многих практических важных ситуациях уровень критичности в поступающих из разных источников сообщениях не задан, так что решение задачи по его установлению и дальнейшим действиям целиком возлагается на систему мониторинга. Системы мониторинга также решают задачи фильтрации (отбрасывания заведомо бесполезных для классификации ситуаций) и дедупликации сообщений (сообщений, поступивших разными путями, возможно от разных элементов инфраструктуры, но свидетельствующих об одной и той же ситуации).

Развитая система способна осуществлять классификацию потока первичных сообщений различной природы посредством применения настроенных правил анализа состояний наблюдаемых компонентов инфраструктуры, правил генерации инцидентов, корреляции или иных доступных аналитических обработчиков. В результате создается поток вторичных событий, в которых уровень критичности классифицированных инцидентов определен (например - авария, предупреждение, информирование, отмена аварии), возможно, точнее исходного с учетом особенностей конкретной инфраструктуры и применения уточняющих методик.

Вторичные события обычно сохраняются в журналах событий системы, а сведения об инцидентах (авариях) выводятся на информационные панели диспетчерских служб, при необходимости, оперативно эскалируются, оповещения направляются ответственным сотрудникам по электронной почте, SMS, автоматически добавляются в системы класса Service Desk, в мессенджеры для запуска сценариев обработки ситуаций.

Ключевая подлежащая решению задача состоит здесь в разработке и применении высокоеффективных и надежных алгоритмов, методов и реализующих их инструментов, которые позволяли бы максимально достоверно и надежно определять уровень важности каждого конкретного события, уровень критичности создаваемого инцидента и инициировать должную обработку ситуации.

Используемые на практике методы действуют готовые сценарии, специальные методы математической статистики, машинного обучения, базы данных управления конфигурацией (CMDB) и ряд других методов. В сложных дина-

мических инфраструктурах требуется учитывать иерархичность, связанность компонентов и их взаимное влияние.

3. «Центральный Пульт»: архитектура платформы, работа с инцидентами и методы определения уровня критичности

3.1. Кратко об архитектуре платформы. Российская высокопроизводительная платформа реализации решений широкого спектра задач мониторинга и управления цифровыми активами «Центральный Пульт» (SAYMON) [4] развивается с 2013 года. Общая архитектура платформы показана на рис. 1.



Рис. 1. Общая архитектура платформы «Центральный Пульт»

Платформа базируется на программно-определенной иерархии наблюдаемых объектов (логических элементов), в качестве которых могут выступать физическое или виртуальное устройство, программный модуль, сервис, технологический или бизнес-показатель и т.д. Графовая модель подразумевает описание объектов в иерархической связности, взаимном влиянии и разграничении прав, обеспечивает высокую скорость работы, возможности автоматизации анализа корневых причин событий (RCA) и мультисервисного использования цифровых массивов.

Встроенное масштабируемое многопользовательское хранилище временных рядов обеспечивает хранение больших объемов высокочастотных данных с требуемыми характеристиками чтения и записи, что позволяет производить быстрый и качественный анализ ситуаций в прошлом и в настоящем, улучшить корреляционный и RCA-анализ, использовать методы машинного обучения для определения аномалий и построения прогнозов.

Взаимодействие пользователей с платформой осуществляется через веб-интерфейс, который доступно и наглядно визуализирует работу наблюдаемых объектов, предоставляет информацию об их состоянии, а также расширенные

возможности управления объектами и связями между ними, событиями, инцидентами, сенсорами, пользователями и их правами и др. REST API предоставляет механизмы для выполнения всех доступных в веб-интерфейсе операций.

3.2. Возможности определения уровня критичности. Набор сенсоров с возможностью кастомизации позволяет гибко настраивать потоки и методы сбора информации. Допустимые обогащение, коррекция форматов данных, конструкторы интерпретации, взаимоувязывания и корреляций, последующих действий помогают реализовать необходимую бизнес-логику. Настраиваемые правила автоматических действий и уведомлений предоставляют возможности выстраивания самовосстанавливающихся систем и приближения к парадигме AIOps.

Первичные события могут поступать в систему от агентов в виде SNMP-ловушек и MQTT-событий (с возможностью предфильтрации), затем ассоциироваться с соответствующими объектами и отображаться в журнале событий в веб-интерфейсе. Привязка событий к объектной модели, правила их классификации осуществляются через графический конструктор или программными сценариями. В отношении событий и классифицированных инцидентов могут быть автоматически выполнены предопределенные операции. Возможно «присоединение» выполнения операции к изменению состояния объекта, что позволяет выстроить автоматическую реакцию системы на происходящие активности.

В целях определения уровня критичности инцидента предусмотрен конструктор условий (графический или программный) для конкретных компонентов наблюдаемой инфраструктуры. Возможно задание степени влияния дочерних узлов описывающей инфраструктуру иерархии на вышестоящие и формирование групповых или синтетических инцидентов. Каждый объект или связь в системе имеет вес. Веса объектов и связей, имеющие общего родителя и находящиеся в одном состоянии, суммируются. Используемый системой алгоритм распространения состояний объектов на вышестоящие узлы показан на рис. 2.

Обоснованное использование обозначенных возможностей требует предварительно решить задачу поиска источников и построения иерархии, а также установить с высокой точностью уровень критичности инцидентов в отношении объектов. Здесь предусмотрены механизмы взаимодействия с CMDB и механизмы автоматического обнаружения (discovery).

4. Заключение

Мониторинг и оперативное управление комплексными цифровыми инфраструктурами - это процесс, предполагающий решение целого спектра взаимно увязанных задач, многие из которых базируются на современном математическом аппарате, алгоритмах и технологиях машинной аналитики данных. Одной из таких задач является достоверная оценка системой мониторинга уровня



Рис. 2. Блок-схема распространения состояний объектов

критичности регистрируемых событий и инициирование должной реакции на классифицированные инциденты.

Реализованные в платформе «Центральный Пульт» механизмы оставляют дополнительные возможности для совершенствования, оценки и анализа развивающихся ситуаций - могут учитываться топологии и иерархии, последовательности, повторяемости событий и измерений. История наблюдений в конкретной инфраструктуре обогащает возможности автоматического анализа оперативно развивающихся ситуаций в будущем. Возможности переноса исторически накопленного опыта от инфраструктуры к инфраструктуре (облако знаний) обеспечивают недостижимый ранее уровень уточнений.

Литература

1. Julian M. Practical Monitoring: Effective Strategies for the Real World. O'Reilly Media, 2018.
2. Turnbull J. The Art of Monitoring. Turnbull Press, 2016.
3. Mauro D., Schmidt K. Essential SNMP. O'Reilly Media, Second Edition, 2005.
4. Программный комплекс «Центральный Пульт» (разработка компании «РОС-СИННО»), <https://cpult.ru>
5. Энсон С. Реагирование на компьютерные инциденты. Прикладной курс. ДМК Пресс, Москва, 2020.
6. Аллакин В. В., Будко Н. П., Васильев Н. В. Общий подход к построению перспективных систем мониторинга распределенных информационно-телекоммуникационных сетей // Системы управления, связи и безопасности. 2021. N4. С. 125–227.

UDC: 004.94

Reliability Of A Redundant Computer System, Taking Into Account The Features Of Information Recovery

V.A. Bogatyrev ^{1,3}, A.V. Bogatyrev ², S.V. Bogatyrev ^{2,3}

¹Department Information Systems Security, Saint-Petersburg State University of
Aerospace Instrumentation, Saint Petersburg, Russia

²Yadro Cloud Storage Development Center, Saint Petersburg, Russia

³ITMO University, Saint Petersburg, Russia

Abstract

Markov models of a fault-tolerant duplicated computer system containing two computers connected to two blocks of two-input memory are proposed. The model is constructed taking into account the physical and subsequent information recovery of memory for various disciplines of information recovery. Comparison of the effectiveness of the disciplines of information recovery of a duplicated computer system is carried out according to a non-stationary readiness coefficient.

The possibilities of increasing the availability of duplicated computer systems are investigated. To carry out these studies, models are proposed to assess the reliability of duplicated computer systems, taking into account the features of physical and informational memory recovery. The models take into account the possible criticality of the system to violations of the continuity of the computing process or to the loss of information after failures of two sides of memory.

The effect on the non-stationary availability factor of the allocation of computing performance resources on the maintenance of functional requests and the recovery of information in memory after its physical recovery is studied.

Keywords: computer system, redundancy, information recovery, availability factor Introduction

1. Introduction

Computer systems operating in real time have high requirements for reliability, fault tolerance [1-4] and timely servicing of requests [5-9]. For fault-tolerant real-time computer systems, in some cases it is necessary to ensure the safety of information, the continuity of the computing process and the timeliness of servicing requests in conditions of failures and deliberate destabilizing influences [10-16]. Support for

timeliness and continuity of the computing process [17] can be based on redundant query service [18-20]. Redundant query service, in addition to improving reliability, timeliness and continuity of calculations, allows you to implement calculation control based on comparison of results. The analyzed feature of the recovery of storage devices is the need for both their physical and subsequent information recovery.

Information recovery involves the use of replicas stored during the operation of the system in workable memory blocks [21]. The loss of all replicas of unique data can lead to the inability to restore information, and to an unrecoverable failure of the computer system. [21]. Information recovery after physical memory recovery based on the use of replicas of up-to-date information contained in a working memory block involves, at a minimum, memory duplication. Reliability models of computer systems containing a computing module and two memory blocks are proposed in [21]. The software models [21] take into account the options with the admissibility and inadmissibility of information recovery after failures of two memory modules. In the first case, it is assumed that the information can be loaded into the physically restored memory from an external source. In the second case, we consider systems for which the information accumulated during operation is unique. After the physical recovery of the memory, information can be entered into it from the working memory. Reliability models of software [21] do not allow to study systems with duplication of memory and processors. The construction of reliability models for duplicated computer systems is of significant interest, since such systems potentially have greater fault tolerance and readiness to organize continuous maintenance. For systems with duplicated memory blocks according to [21], continuity of service is lost during the failure and recovery of a single computer or two memory blocks. For systems with duplicated memory blocks and calculators, continuity of service is lost during the failure of two calculators or two memory blocks. Splitting the load into two calculators allows you to reduce the average time spent by requests and increase the likelihood of timely servicing them, which is important for real-time systems. Duplication of calculators makes it possible to increase the reliability and reliability of service when executing queries in two calculators with further comparison of results, and if they do not match, repeat calculations. When the input stream is heterogeneous in terms of the allowable waiting time for requests, reserving requests with the smallest waiting time margin also increases the probability of timely execution of the most critical to waiting requests [21].

The purpose of this article is to study the possibilities of increasing the availability of duplicated computer systems containing two computers connected to two blocks of two-input memory. To conduct research, models for assessing the reliability of duplicated computer systems should be developed, taking into account the features of physical and informational memory recovery. Models should take into account the

possible criticality of the system to disruptions of the continuity of the computing process or to the loss of information after failures of two sides of memory.

The models being developed should take into account the allocation of computing resources for servicing the request flow and for informational memory recovery carried out with the participation of the computer after physical memory recovery.

2. Construction of reliability models for physical and informational memory recovery

As an object of research, a duplicated computer system is considered, including two computer nodes with two blocks of two-input memory pumped to them). The analyzed fault-tolerant redundant computer system provides for the implementation of duplicated calculations with replication of results in two available memory blocks.

The structure of the duplicated computing system under consideration is shown in Fig. 1. Two-input memory (M) allows it to be accessed from calculators (C) of different nodes, which makes it possible, in case of degradation, to implement calculations on the equipment of two nodes stored after failures.

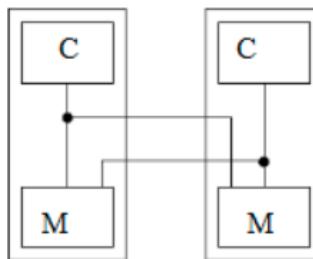


Fig. 1. The structure of the duplicated computer system

The system implements two-stage memory recovery. At the first stage, the physical recovery of memory is carried out, and at the second – its informational recovery. Let's consider three options for organizing the recovery of information and the system as a whole, taking into account various requirements for ensuring the continuity of the computing process and the safety of information after the accumulation of failures.

For the first option, let's assume that ensuring the continuity of the computing process is not required, and the information being restored is not unique and can be entered into physically restored memory from some source.

For the second option, the recovered information is unique and can be loaded into physically recovered memory only based on replicas stored in the failed storage device. If two storage devices fail, the information cannot be restored to a physically

restored storage device, which can lead to the inability of the system to function, that is, to its failure (transition to a state of non-recoverable failure).

For the third option, interrupts of the computing process are not allowed. The system goes into a non-recoverable state when two calculators or two memory blocks fail.

Markov reliability models for the first, second and third variants of the organization of computer system maintenance are represented by diagrams of states and transitions according to Fig. 2, Fig. 3. and Fig.4.

We encode the states with a matrix, the upper line of which displays the states of two computers, and the lower one - two memory blocks. The operability of the blocks is displayed as "1", and the failure as "0". If the memory block is physically restored, but the necessary information is not entered into it, then the state is indicated as - "F". States are invariant to block numbering.

The probabilities of states 1, 2, ... of the system are denoted as P_1, P_2, \dots . In the diagrams according to Fig. 2 - Fig. 4, the failure rates and recoveries of computing blocks Λ_1 and μ_1 , the failure rates of memory blocks Λ_2 , as well as the intensity of their physical and informational recovery are indicated. The values of the intensity of information recovery μ_{22} correspond to the condition that all the resources of one of the calculators are directed to restoring the information of one memory block.

The intensity of information recovery depends on the shares of stored performance resources of efficient computers allocated for servicing the flow of functional requests and for information recovery of memory. Let's denote by β the share of performance resources of one computer allocated for information recovery of memory. The intensity of the transition from state "9" to state "2" is defined as $\beta\mu_{22}$. It should be noted that the shares of β resources allocated for informational memory recovery affect the operability of system states. So the state "9" at $\beta = 1$ is not operable, since the only workable computer is used exclusively for information recovery, in which the service of functional requests is not performed. With $\beta < 1$, the state "9" is operable, and the probability of timely servicing requests in this state depends on β .

The probabilities of the system state are denoted as P_1, P_2, \dots . The presented diagrams indicate the failure rates and recoveries of the computing block Λ_1 and μ_1 , the failure rates of memory blocks Λ_2 , and the intensity of their physical and information recovery μ_{21} and μ_{22} .

The presented diagrams of states and transitions allow you to create a system of algebraic or differential equations, the solution of which allows you to determine the probabilities of all states of the system. As a result, it is possible to determine the dependence of the probability of system operability on the time of its operation, as well as the stationary and non-stationary availability coefficient for restored systems.

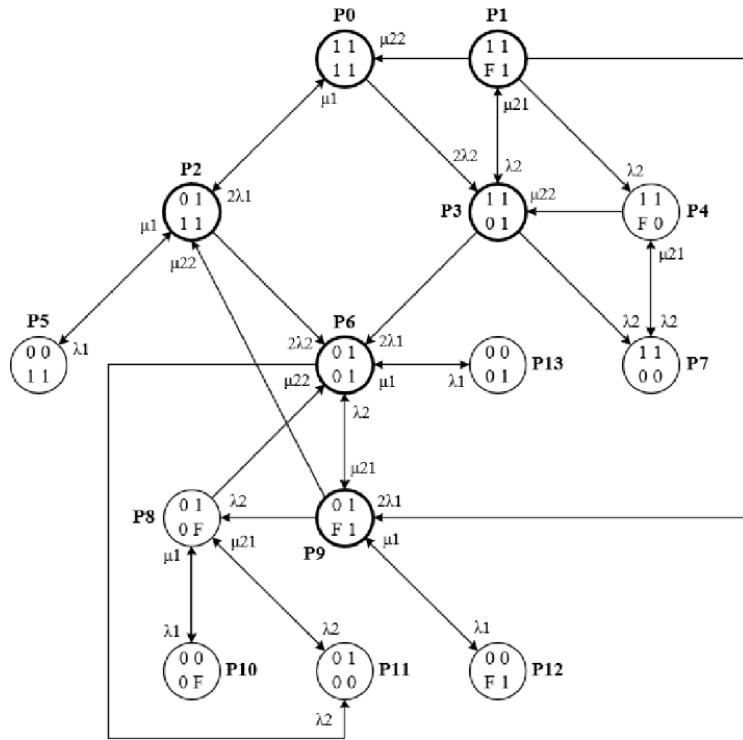


Fig. 2. Markov model of a duplicated system for physical and informational recovery of duplicated memory, with the possible recovery of information after loss (violation of the continuity of calculations during recovery is permissible)

For example, for the diagram of states and transitions in Fig.1, the system of differential equations presented for solution in the Mathcad 15 computer mathematics system has the form:

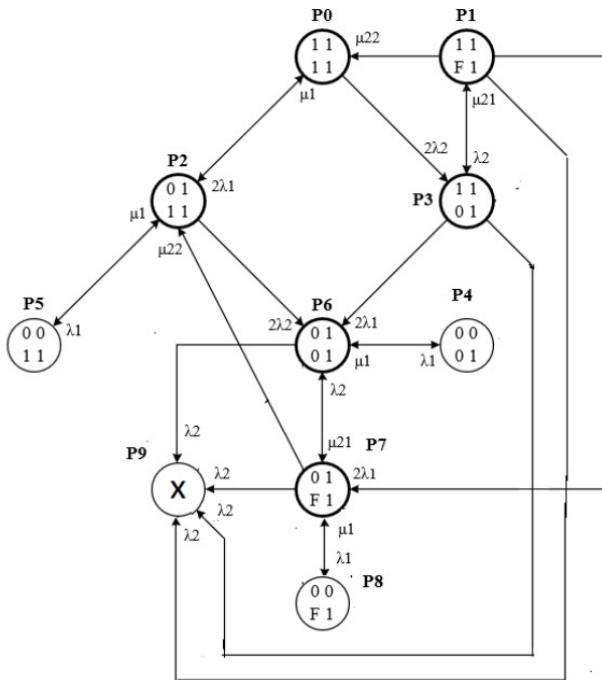


Fig. 3. Markov model of a duplicated computer system for physical and informational recovery of duplicated memory when information loss is unacceptable

$$D(t, p) := \begin{bmatrix} -(2\lambda_1 + 2\lambda_2)p_0 + \mu_1 \cdot p_2 + \mu_{22} \cdot p_1 \\ -(2\lambda_1 + \lambda_2 + \lambda_2 + \mu_{22})p_1 + \mu_{21} \cdot p_3 \\ -(2\lambda_2 + \lambda_1 + \mu_1)p_2 + 2\lambda_1 \cdot p_0 + \mu_{22} \cdot p_9 + \mu_1 \cdot p_5 \\ -(2\lambda_1 + \lambda_2 + \mu_{21})p_3 + \lambda_2 \cdot p_1 + \mu_{22} \cdot p_4 + 2\lambda_2 \cdot p_0 \\ -(\lambda_2 + \mu_{22})p_4 + \lambda_2 \cdot p_1 + \mu_{21} \cdot p_7 \\ -(\mu_1)p_5 + \lambda_1 \cdot p_2 \\ -(\mu_{21} + \lambda_2 + \lambda_1)p_6 + 2\lambda_2 \cdot p_2 + \lambda_2 \cdot p_9 + \mu_{22} \cdot p_8 + 2\lambda_1 \cdot p_3 + \mu_1 \cdot p_{13} \\ -(\mu_{21})p_7 + \lambda_2 \cdot p_3 + \lambda_2 \cdot p_4 \\ -(\mu_{22} + \lambda_1 + \lambda_2) \cdot p_8 + \lambda_2 \cdot p_9 + \mu_1 \cdot p_{10} + \mu_{21} \cdot p_{11} \\ -(\mu_{22} + \lambda_2 + \lambda_2 + \lambda_1) \cdot p_9 + \mu_1 \cdot p_{12} + \mu_{21} \cdot p_6 + 2\lambda_1 \cdot p_1 \\ -(\mu_1) \cdot p_{10} + \lambda_1 \cdot p_8 \\ -(\mu_{21}) \cdot p_{11} + \lambda_2 \cdot p_8 + \lambda_2 \cdot p_6 \\ -(\mu_1) \cdot p_{12} + \lambda_1 \cdot p_9 \\ -(\mu_1) \cdot p_{13} + \lambda_1 \cdot p_6 \end{bmatrix} \quad p := \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Z := rkfixed(p, 0, 1000, 10000, D) n := 0..10000

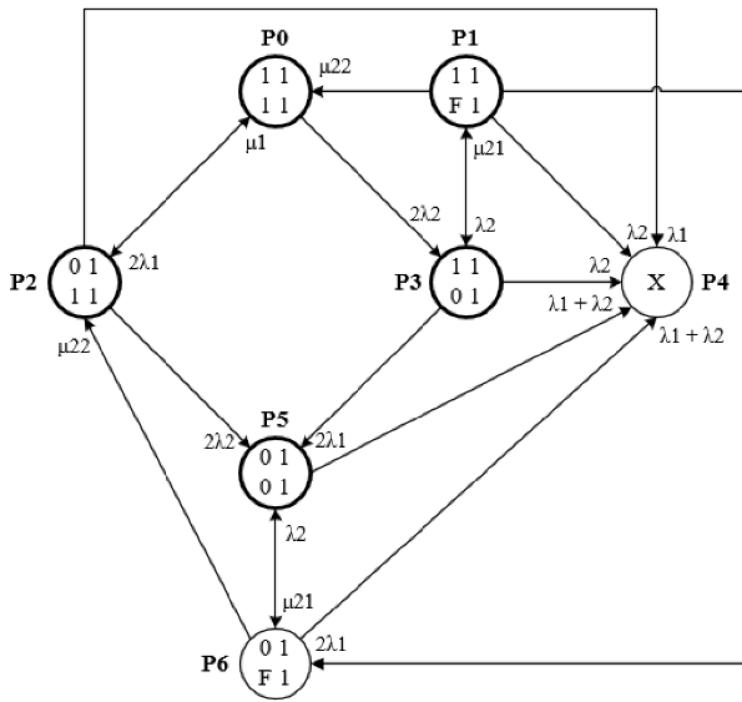


Fig. 4. Markov model of a duplicated computer system with the inadmissibility of interrupting the computing process

The expression $D(t,p)$ corresponds to the matrix form of specifying a system of differential equations in Mathcad 15, the vector (p) sets the initial state of the system.

3. Example of calculating the reliability of a computer system

The results of the calculation of the non-stationary availability coefficient of duplicated computer systems, taking into account and without taking into account the requirement to ensure the continuity of the computing process, are shown in Fig. 5 a and b.

The calculation was performed at $\Lambda_1 = 10^{-4} \text{ 1/h}$, $\Lambda_2 = 210^{-4} \text{ 1/h}$, $\mu_1 = 1 \text{ 1/h}$, $\mu_{21} = 1 \text{ 1/h}$, $\mu_{22} = 11 \text{ 1/h}$.

The presented dependencies confirm the significance of the influence of the factors under consideration on the reliability of the computer systems under study.

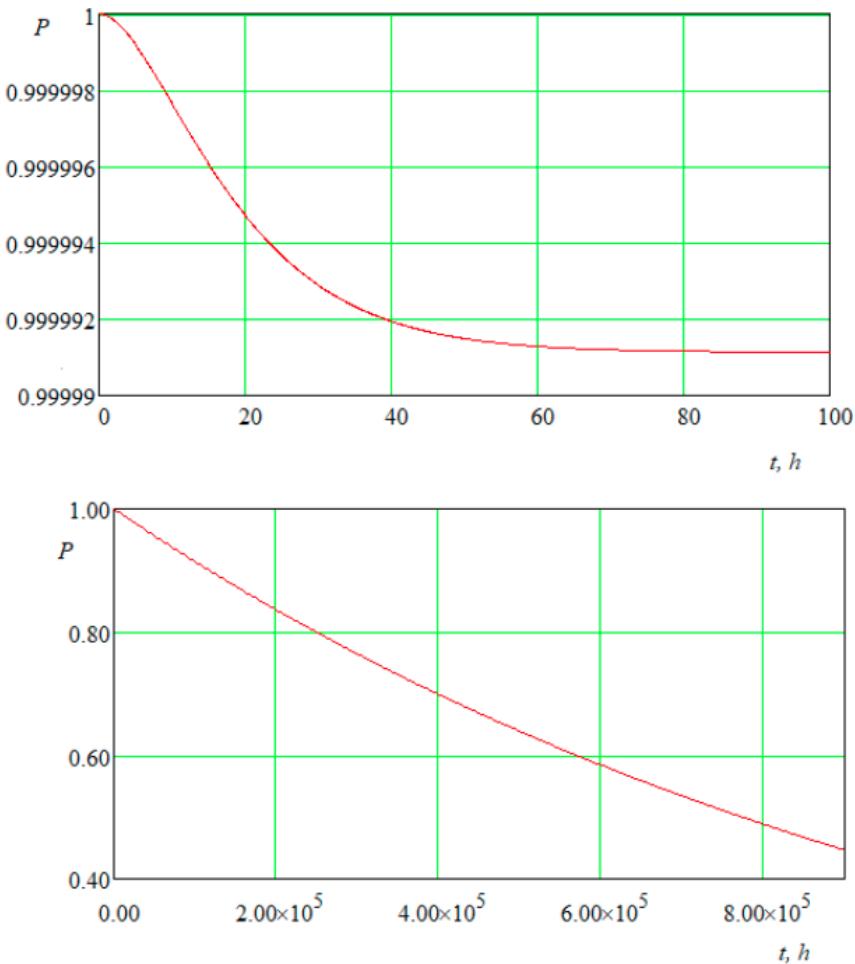


Fig. 5. The coefficient of unsteady availability of duplicated computer systems without condition (a) and with condition (b) ensuring the continuity of the computing process.

4. Assessment of the impact of combining information recovery of memory and maintenance of functional requests on the readiness of a duplicated computer system

Let's analyze the effect on the coefficient of non-stationary readiness of the allocation of computing resources for servicing the flow of requests and for information recovery of memory. Consider duplicated systems that allow information recovery of memory after failures of two memory blocks. The diagram of the states and transitions of the system under study is shown in Fig. 2. The impact on the non-stationary availability coefficient of the distribution of computing performance

resources on the actual calculations and information recovery of memory is significant for the state "9". This is due to the fact that in the state "9" there is one workable computer, one workable memory block and one memory block that is at the stage of information recovery, which is carried out with the participation of the computer.

The calculation results are shown in Fig. 6, in which curves 1-4 correspond to the value $\beta = 0.95, 0.5, 0.2$ and 0.1 . From Fig. 6 it can be seen that the readiness of the duplicated system increases when allocating more computing performance resources for information memory recovery than for query servicing. However, an increase in the share of computing resources involved in loading memory after its physical recovery negatively affects the likelihood of timely execution of requests. Thus, it is necessary to find a compromise to resolve the noted technical contradiction. As a result, there is a task and optimization of the share of computing resources allocated for calculations and for information recovery of memory.

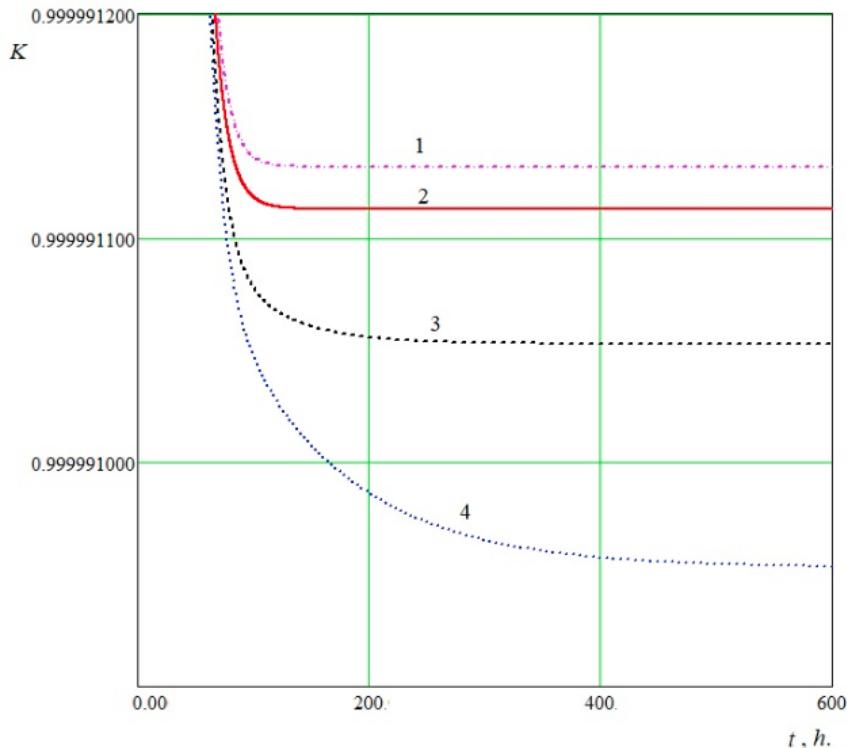


Fig. 6. The effect on the non-stationary availability coefficient of the computing performance distribution on query execution and information memory recovery

5. Conclusion

A Markov model of a fault-tolerant duplicated computer system is proposed, reflecting the stages of physical and informational memory recovery.

The proposed model takes into account the features of physical and informational memory recovery with the possible criticality of the system to the continuity of the computing process or to the loss as a result of failures of all created replicas of unique information generated during the operation of the system.

The model takes into account the possibility of allocating computing performance resources for solving functional problems and restoring information.

The effect on the unsteady availability coefficient of the resource allocation of computing performance on the maintenance of functional requests and the recovery of information in memory after its physical recovery is shown.

REFERENCES

1. Aysan H. Fault-tolerance strategies and probabilistic guarantees for real-time systems Mälardalen University, Västerås, Sweden. 2012. 190 p.
2. Merindol P. Improving Load Balancing with Multipath Routing / P. Merindol, J. Pansiot, S. Cateloin // Proc. of the 17-th International Conference on Computer Communications and Networks, IEEE ICCCN 2008. – 2008. – P. 54-61
3. Chen W.H. and Tsai J.C. (2014) Fault-Tolerance Implementation in Typical Distributed Stream Processing Systems.
4. Machida F., Kawato M., Maeno Y. Redundant virtual machine placement for fault-tolerant consolidated server clusters. In: IEEE Network Operations and Management Symposium, pp. 32–39. IEEE Press, Osaka (2010), doi: 10.1109/NOMS.2010.5488431.02071-85
5. Koren I.: Fault tolerant systems. Morgan Kaufmann publications, San Francisco 2009 378 p.
6. Shooman M.L., Reliability of computer systems and networks. John Wiley Sons Inc., 2002.
7. Bennis M., Debbah M. Poor H.V. Ultrareliable and Low-Latency Wireless Communication: Tail, Risk and Scale. Proc. IEEE 2018, 106, 1834–1853. DOI: 10.1109/JPROC.2018.2867029
8. Kim I S., Choi Y., Constraint-aware VM placement in heterogeneous computing clusters. //Cluster Comput. 23, 71–85 (2020). <https://doi.org/10.1007/s10586-019-02966-6>.
9. Sachs J., G. Wikström G., Dudda T., Baldemair R. Kittichokechai K., 5G Radio Network Design for Ultra-Reliable Low-Latency Communication.// IEEE Netw. 2018, 32, 24–31. DOI:10.1109/MNET.2018.1700232.

10. Ji H. Park S , Yeo J, Kim Y., LeeJ., Shim B., Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects.// IEEE Wirel. Commun. 2018, 25, 124–130. DOI:10.1109/MWC.2018.1700294.
11. Samarasinghe S., Neural Networks for Applied Sciences and engineering: from Fundamentals to Complex Pattern Recognition /– Boston: Auerbach publications, 2016. – 570 p.
12. Siddiqi M., H. Yu H., Joung j., 5G Ultra-Reliable Low-Latency Communication Implementation Challenges and Operational Issues with IoT Devices //Electronics 2019, 8, 981; doi:10.3390/electronics8090981 www.mdpi.com/journal/electronics. DOI: 10.3390/electronics8090981.
13. Zakoldaev D.A., Korobeynikov A.G, Zharinov I.O. , Zharinov, O.O. : Industry 4.0 vs Industry 3.0: the role of personnel in production//IOP Conference Series: Materials Science and Engineering, 2020, Vol. 734, No. 1, pp. 012048
14. Malik V., Barde C.R. Live migration of virtual machines in cloud environment using prediction of CPU usage // International Journal of Computer Applications. 2015. V. 117 N 23. P. 1–5. doi: 10.5120/20691-3604
15. Rausand M., Hoyland A., System reliability theory. John Wiley Sons Inc., 2004. . Greenan M.K, Plank J. S., Wylie J. J. Mean time to meaningless: MTTDL. Markovmodels, and storage system reliability, HotStorage (2010).
16. Tatarnikova T.M., Poyanova, E.D.: Differentiated capacity extension method for system of data storage with multilevel structure. //Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2020, 20(1), . 66–73 DOI: 10.17586/2226-1494-2020-20-1-66-73
17. Astakhova T. N., Verzun N. A., Kasatkin V. V., Kolbanev M. O., Shamin A. A.: Sensor network connectivity models. Informatsionno-upravliaushchie sistemy // [Information and Control Systems], 2019, no. 5, pp. 38–50. doi:10.31799/16848853-2019-5-38-50. <https://doi.org/10.31799/1684-8853-2019-5-38-50>.
18. Bogatyrev V.A., Derkach A.N. Evaluation of a Cyber-Physical Computing System with Migration of Virtual Machines during Continuous Computing // Computers - 2020, Vol. 9, No. 2, pp. 42. DOI 10.3390/computers9020042.
19. Bogatyrev V.A. , Bogatyrev A.V. , Bogatyrev S.V. : Redundant Servicing of a Flow of Heterogeneous Requests Critical to the Total Waiting Time During the Multi-path Passage of a Sequence of Info-Communication Nodes. Lecture Notes in Computer Science// Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020. Vol. 12563. pp. 100-112. DOI 10.1007/978-3-030-66471-8_9
20. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V. Redundant multi-path service of a flow heterogeneous in delay criticality with defined node passage paths //

- Journal of Physics: Conference Series, Volume 1864, 13th Multiconference on Control Problems (MCCP 2020) 6-8 October 2020, Saint Petersburg, Russiaal 2021 J. Phys.: Conf. Ser. 1864 012094 - 2021, Vol. 1864, 012094, No. 1, pp. 012094 . DOI 10.1088/1742-6596/1864/1/012094.
21. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Reliability of Computer Systems during Physical and Informational Recovery of Duplicated Memory. CEUR Workshop Proceedings. 2021. Vol. 3057. pp. 274-279.

UDC: 519.216

Non-asymptotic Confidence Estimation of the Autoregressive Parameter in ARMA(1,q) Model

S.E. Vorobeychikov¹ and A.V. Pupkov²

^{1,2}Tomsk State University, Lenin Av. 36 , Tomsk, Russia

sev@mail.tsu.ru, andrewpupkov@gmail.com

Abstract

The article deals with the problem of confidence estimation of the autoregressive parameter of the ARMA(1, q) process in a non-asymptotic statement. It is assumed that the noise of the process is Gaussian with an unknown variance. The problem is solved by using a sequential modification of the Yule-Walker estimate.

Keywords: ARMA, non-asymptotic estimation, Yule–Walker estimator, martingales.

1. Introduction

Autoregressive-moving-average (ARMA) models are often used in many engineering applications. For example, in speech recognition [1], finance [2], in the area of modeling of stochastic processes with prescribed distribution [3] etc. There are many results concerning the problem of estimating parameter problems in ARMA models and autoregressive processes with noise [4, 5, 6]. Many results are obtained in an asymptotic setting, but the question of the quality of estimation for small and moderate sample sizes is open. Later the sequential estimation methods were developed in [7, 8, 9, 10] to investigate the non-asymptotic characteristics of the estimators of the parameters. In this paper we present the results of the confidence estimation of the parameter AR(1) in the ARMA(1, q) process using sequential approach .

The paper is organized as follows. Section 2 describes the process under study. In section 3 we briefly present some auxiliary results. In section 4, a sequential estimator is introduced, which is used in section 5 in the construction of a confidence interval.

2. Problem statement

Consider ARMA(1, q) process

$$x_k = \theta x_{k-1} + \eta_k, \quad \eta_k = \xi'_k \lambda, \quad k = 1, 2, \dots, \quad (1)$$

where $|\theta| < 1$ is an unknown parameter, $\xi_k = (\varepsilon_k, \dots, \varepsilon_{k-q})'$, $\{\varepsilon_k\}$ is a sequence of independent Gaussian random variables with zero mean $\mathbf{E}\varepsilon_k = 0$ and variance $\mathbf{E}\varepsilon_k^2 = b^2$, $\lambda = (1, \lambda_1, \dots, \lambda_q)'$ is a vector of parameters of the moving average part; the prime ('') stands for transposition. It's assumed that value b and vector λ are unknown subject to $\lambda' \lambda \leq \Lambda < +\infty$. Note that the sequence $\{\varepsilon_k\}$ doesn't depend on the random variable x_0 .

The aim is to construct the confidence interval for parameter θ by using observations $\{x_k\}$.

3. Estimation of the variance of the MA(q) part

Procedure of the confidence estimation of the parameter θ includes three stages. On the first stage we introduce pilot estimator of θ by sample of fixed size. On the second stage we estimate variance of the noise η_k (moving average part) to eliminate the influence of unknown factor b in the noise of the process. Then, on the third stage, we introduce system of sequential estimates of the parameter θ to obtain the confidence interval of θ .

First initially we construct a pilot Yule–Walker estimator of the parameter θ

$$\tilde{\theta}_{n_1} = \left(\sum_{k=q+1}^{n_1} x_{k-q-1} x_{k-1} \right)^{-1} \sum_{k=q+1}^{n_1} x_{k-q-1} x_k, \quad n_1 \geq q+1 \quad (2)$$

by sample size of $n_1 - q$. Then we get an estimate of the variance η_k in the form

$$\Gamma_{n_1, n_2} = C_{n_2} S_{n_1, n_2}, \quad (3)$$

where

$$C_{n_2} = (n_2 - n_1 - 2)^{-1}, \quad S_{n_1, n_2} = \sum_{k=n_1+1}^{n_2} \left(x_k - \tilde{\theta}_{n_1} x_{k-1} \right)^2. \quad (4)$$

An important result for construction a confidence interval contains

Lemma 1. Let $G(y)$ be the distribution function of a random variable $\Gamma_{n_1, n_2} / \varkappa_\lambda^2$, where $\varkappa_\lambda^2 = \mathbf{E}\eta_k^2 = b^2 \lambda' \lambda$ is the variance of η_k . Assuming that $\lambda' \lambda \leq \Lambda < +\infty$, the inequality holds true

$$G(y) = P \left(\frac{\Gamma_{n_1, n_2}}{\varkappa_\lambda^2} < y \right) \leq P \left(\sum_{k=n_1+1}^{n_2} \left(\frac{\varepsilon_k}{b} \right)^2 < \frac{y\Lambda}{C_{n_2}} \right). \quad (5)$$

4. Sequential Yule-Walker estimator

Now we construct a sequential modification of the estimator (2). Using the approach proposed in [10], we divide the set of indexes of the process observations into a $q + 1$ subsets in the form

$$T(n) = \sum_{i=1}^{q+1} T_i(n), \quad T_i(n) = \{k : k = n_2 + q + i + (q + 1)j, j = 0, 1, \dots; k \leq n\}. \quad (6)$$

For any $h > 0$, we introduce a system of stopping moments

$$\tau_i(h) = \inf \left\{ n > n_2 + q : \sum_{k=n_2+q+1}^n \chi_{\{k \in T_i(n)\}} \frac{x_{k-q-1}^2}{\Gamma_{n_1, n_2}} \geq h \right\}, \quad i = \overline{1, q+1}, \quad (7)$$

where $\chi_{\{A\}}$ is indicator function of the event A ; parameter h defines the accuracy of the estimator θ .

Construct a system of sequential estimates for each of the subsets of indexes $T_i(\tau_i(h))$ in the form

$$\hat{\theta}^{(i)}(h) = \frac{\vartheta^{(i)}(h)}{s^{(i)}(h)}, \quad i = \overline{1, q+1}, \quad (8)$$

$$\vartheta^{(i)}(h) = \sum_{k=n_2+q+1}^{\tau_i(h)} D_i(k, h) x_k, \quad s^{(i)}(h) = \sum_{k=n_2+q+1}^{\tau_i(h)} D_i(k, h) x_{k-1}, \quad (9)$$

$$D_i(k, h) = \chi_{\{k \in T_i(k)\}} x_{k-q-1} \sqrt{\frac{\beta_k^{(i)}(h)}{\Gamma_{n_1, n_2}}} \quad (10)$$

where

$$\beta_k^{(i)}(h) = \begin{cases} 1, & \text{if } n_2 + q < k < \tau_i(h); \\ \alpha_i(h), & \text{if } k = \tau_i(h); \end{cases} \quad (11)$$

and $0 < \alpha_i(h) \leq 1$ is the coefficient uniquely determined from the equation

$$\sum_{k=n_2+q+1}^{\tau_i(h)-1} \chi_{\{k \in T_i(n)\}} \frac{x_{k-q-1}^2}{\Gamma_{n_1, n_2}} + \alpha_i(h) \frac{x_{\tau_i(h)-q-1}^2}{\Gamma_{n_1, n_2}} = h. \quad (12)$$

Next, investigate the properties of estimate deviation $\hat{\theta}^{(i)}(h) - \theta$. Substituting the equation of the process (1) in (8) we get

$$\hat{\theta}^{(i)}(h) - \theta = \varkappa_\lambda \frac{\tilde{\zeta}^{(i)}(h)}{s^{(i)}(h)} \quad (13)$$

where

$$\tilde{\zeta}^{(i)}(h) = \sum_{k=n_2+q+1}^{\tau_i(h)} D_i(k, h) \tilde{\eta}_k, \quad \tilde{\eta}_k = \varkappa_\lambda^{-1} \eta_k. \quad (14)$$

The following result plays a key role in constructing the confidence interval. Formulate it in the form of the lemma.

Lemma 2. Let $\{\varepsilon_k\}$ in the process (1) be a sequence of independent Gaussian variables with zero means $\mathbf{E}\varepsilon_k = 0$ and variances $\mathbf{E}\varepsilon_k^2 = b^2$. Then for any $h > 0$ and any $i = \overline{1, q+1}$, the variable $h^{-1/2}\tilde{\zeta}^{(i)}(h)$ has a standard Gaussian distribution, i.e. $\text{Law}(h^{-1/2}\tilde{\zeta}^{(i)}(h)) = \mathcal{N}(0, 1)$.

Proof. The result of Lemma 2 naturally follows from the fact that sequences $(M_n^{(i)}, \mathcal{F}_n^{(i)})_{n>n_2+q}$ are martingales, as well as from the definition of stopping times (7) and from the Theorem 1 in [7]. Here

$$M_n^{(i)} = \sum_{k=n_2+q+1}^n \chi_{\{k \in T_i(k)\}} \frac{x_{k-q-1} \tilde{\eta}_k}{\sqrt{\Gamma_{n_1, n_2}}}, \quad (15)$$

$$\mathcal{F}_n^{(i)} = \sigma \{x_0, \varepsilon_1, \dots, \varepsilon_{d_i(n)}\}, \quad d_i(n) = \max \{k : k \in T_i(n)\}, \quad i = \overline{1, q+1}. \quad (16)$$

■

In the next section we will provide the main result.

5. Non-asymptotic confidence interval

Theorem 1. Let in the ARMA(1, q) process, described by the expression (1), $\{\varepsilon_k\}$ be a sequence of independent Gaussian random variables with zero mean $\mathbf{E}\varepsilon_k = 0$ and variance $\mathbf{E}\varepsilon_k^2 = b^2$ and sequential point estimators are defined by (7), (8) and (11). Then for any $h > 0$ the inequality

$$\begin{aligned} \inf_{|\theta|<1} P \left(\frac{1}{h\sqrt{\Gamma_{n_1, n_2}}} \left| \frac{s_*(h)}{q+1} \sum_{i=1}^{q+1} (\hat{\theta}^{(i)}(h) - \theta) \right| < z \right) &\geq \\ &\geq (q+1) \int_0^{+\infty} \left[2\Phi \left(z \sqrt{\frac{yhC_{n_2}}{\Lambda}} \right) - 1 \right] \frac{y^{N/2-1} e^{-y/2}}{2^{N/2} \Gamma(N/2)} dy - q, \end{aligned} \quad (17)$$

holds true. Here $s_*(h) = \min_{k=\overline{1, q+1}} |s^{(k)}(h)|$, $\Gamma(x)$ is Gamma-function, $N = n_2 - n_1$ and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy, \quad x \geq 0. \quad (18)$$

Proof. Using the inequality $|s_*(h)/s^{(k)}(h)| \leq 1$ for any $k = 1, \dots, q+1$ we find the lower bound of the following probability

$$\begin{aligned} P\left(\frac{|s_*(h)|}{h(q+1)\sqrt{\Gamma_{n_1,n_2}}}\left|\sum_{i=1}^{q+1}(\hat{\theta}^{(i)}(h)-\theta)\right| < z\right) &\geq \\ \geq P\left(\frac{1}{\sqrt{h\Gamma_{n_1,n_2}}}\sum_{i=1}^{q+1}\left|\tilde{\zeta}^{(i)}(h)\right| < \frac{z(q+1)\sqrt{h}}{\varkappa_\lambda}\right) &\geq \\ \geq (q+1)\mathbf{E}P\left(\frac{1}{\sqrt{h}}\left|\tilde{\zeta}^{(1)}(h)\right| < \frac{z\sqrt{h\Gamma_{n_1,n_2}}}{\varkappa_\lambda}\middle|\mathcal{F}_{n_2}\right) - q. & \quad (19) \end{aligned}$$

here $\mathcal{F}_{n_2} = \sigma\{x_0, \varepsilon_1, \dots, \varepsilon_{n_2}\}$. Based on the result of the Lemma 1, we get

$$\begin{aligned} \mathbf{E}P\left(\frac{1}{\sqrt{h}}\left|\tilde{\zeta}^{(1)}(h)\right| < \frac{z\sqrt{h\Gamma_{n_1,n_2}}}{\varkappa_\lambda}\middle|\mathcal{F}_{n_2}\right) &= \\ = \int_0^{+\infty}P\left(\frac{1}{\sqrt{h}}\left|\tilde{\zeta}^{(1)}(h)\right| < z\sqrt{hy}\middle|\mathcal{F}_{n_2}\right)dG(y) &= \\ = 1 - \int_0^{+\infty}G(y)dP\left(\frac{1}{\sqrt{h}}\left|\tilde{\zeta}^{(1)}(h)\right| < z\sqrt{hy}\middle|\mathcal{F}_{n_2}\right) &\geq \\ \geq \int_0^{+\infty}P\left(\frac{1}{\sqrt{h}}\left|\tilde{\zeta}^{(1)}(h)\right| < z\sqrt{hy}\middle|\mathcal{F}_{n_2}\right)dP\left(\sum_{k=n_1+1}^{n_2}\left(\frac{\varepsilon_k}{b}\right)^2 < \frac{y\Lambda}{C_{n_2}}\right). & \quad (20) \end{aligned}$$

Combining results (19) and (20), and also noting that, the quantity $h^{-1/2}\tilde{\zeta}^{(1)}(h)$ has a standard Gaussian distribution, and $b^{-2}(\varepsilon_{n_1+1}^2 + \dots + \varepsilon_{n_2}^2)$ has a chi-squared distribution with $N = n_2 - n_1$ degrees of freedom, we come to the result of Theorem 1. ■

Remark 1. Similar to Proposition 1 in [10], it can be shown that value $s_*(h)/h$ converges to a constant almost surely.

6. Conclusion

The paper proposes an non-asymptotic method of confidence estimation of the autoregressive parameter of the ARMA(1, q) process using a sequential modification of the Yule-Walker estimate. The results can be used in identification and control problems.

REFERENCES

1. S. Ganapathy, Robust speech processing using ARMA spectrogram models, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5029–5033. doi:10.1109/ICASSP.2015.7178928.
2. N. Kuznietsova, P. Bidyuk, Heteroskedasticity models for financial processes modelling and forecasting, in: 2020 IEEE Third International Conference on Data Stream Mining Processing (DSMP), 2020, pp. 310–315. doi:10.1109/DSMP47368.2020.9204313.
3. S. V. Garbar, Modification of ARMA (1, 1) model to simulate strictly stationary random series with uniform distribution, *Journal of Physics: Conference Series* 1352 (1) (2019) 012020. doi:10.1088/1742-6596/1352/1/012020.
URL <https://doi.org/10.1088/1742-6596/1352/1/012020>
4. M. K. Hasan, A. K. M. Z. R. Chowdhury, M. R. Khan, Identification of autoregressive signals in colored noise using damped sinusoidal model, *IEEE Transactions on Circuits and Systems -I: Fundamental Theory and Applications* 50 (2003) 1475–1480.
5. C. Wei, Y. Hao, K. Ching, Interval estimation for a first-order positive autoregressive process, *Journal of Time Series Analysis* 39 (2018) 1475–1480. doi:10.1111/jtsa.12297.
6. A. Mahmoudi, M. Karimi, H. Amindavar, Parameter estimation of autoregressive signals in presence of colored ar(1) noise as a quadratic eigenvalue problem, *IEEE Transactions on Circuits and Systems -I: Fundamental Theory and Applications* 92 (2012) 1151–1156.
7. V. Konev, On one property of martingales with conditionally gaussian increments and its application in the theory of non-asymptotic inference, *Doklady Mathematics* 94 (2016) 676–680. doi:10.1134/S1064562416060235.
8. S. Vorobeychikov, Y. Burkatskaya, Non-asymptotic confidence estimation of the autoregressive parameter in ar(1) process with an unknown noise variance, *Austrian Journal of Statistics* 49 (2020) 19–26. doi:10.17713/ajs.v49i4.1121.
9. V. Konev, B. Nazarenko, Sequential fixed accuracy estimation for nonstationary autoregressive processes, *Annals of the Institute of Statistical Mathematics* 72 (1) (2020) 235–264. doi:10.1007/s10463-018-0689-2.
URL <https://link.springer.com/article/10.1007/s10463-018-0689-2>
10. V. Konev, A. Pupkov, Confidence estimation of autoregressive parameters based on noisy data, *Automation and Remote Control* 82 (2021) 1030–1048. doi:10.1134/S0005117921060059.

UDC: 004.94

Cluster With Functional Heterogeneity Of Nodes With Requests Of Different Criticality to Delays

V.A. Bogatyrev ^{1,3}, A.V. Bogatyrev ², S.V. Bogatyrev ^{2,3}

¹Department Information Systems Security, Saint-Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia

²Yadro Cloud Storage Development Center, Saint Petersburg, Russia

³ITMO University, Saint Petersburg, Russia

Abstract

Analytical models and methods for ensuring functional reliability of clusters based on redundant query service are considered. The analysis of fault-tolerant real-time cluster systems becomes more complicated when the functional configuration of cluster nodes is heterogeneous and when the input stream is heterogeneous in terms of the functionality of requests and their criticality to service delays. The nodes of the cluster under consideration combine a general-purpose server and functionally oriented servers. Cluster nodes can be equipped with different sets of functionally oriented servers. In the cluster systems under consideration, the request is first served in a general-purpose server, and then in a functionally oriented server. Waiting in the queue of a general-purpose server can significantly affect the overall delay of two-stage query servicing in the cluster, and therefore these delays should be taken into account.

The article investigates the possibility of increasing the probability of timely execution in a cluster of a heterogeneous in functionality and acceptable waiting time for a stream of requests. The assessment of timeliness is carried out taking into account the accumulation of waiting delays in queues at the service stages in a general-purpose server and in one of the functionally oriented servers. The expediency of reserving the most delay-critical functional requests is evaluated. The existence of an optimal multiplicity of reservation requests is shown

Keywords: probability of timely maintenance, cluster, functional heterogeneity, criticality of requests to delays, maximum allowable waiting, real time

1. Introduction

Distributed computer systems are subject to high requirements in terms of reliability, performance [1-6] and security, achieved by consolidating resources and combining them into clusters [6 11]. Currently, there are many works on the study

of methods for improving and evaluating the structural reliability of cluster systems [12-16]. For real-time cluster systems, the key is to ensure functional reliability, which is understood as the reliable functioning of the system, provided the timely servicing of requests. The tasks of evaluating and ensuring the functional reliability of distributed systems have not yet been fully investigated.

This article discusses models and methods for ensuring the functional reliability of distributed systems when executing queries in clusters with functional heterogeneity of nodes. At the same time, the possibility of increasing functional reliability on the basis of redundant real-time maintenance of a heterogeneous flow of requests in terms of functionality is being considered.

The methodology of redundant service is the development of the concepts of multipath routing], multicast transmissions, broadcast service and dynamic distribution of requests [17,18]. The discipline of redundant service in multichannel systems is known [18], which provides for the creation and parallel execution of copies of requests, provided they are received when there are free service channels. The discipline of [18] makes it possible to increase the probability of timely maintenance only in the absence of queues at nodes and the presence of unoccupied channels [18]. For computer systems of cluster architecture with a heterogeneous flow of requests over the allowable waiting time, the probability of timely servicing of the most delay-critical requests can be achieved as a result of traffic prioritization. Traffic prioritization can be combined with redundant query service [19]. This combination is especially effective when there is a high probability of transmission errors, including under the influence of destabilizing factors [19].

The combination of reserved and priority services can be effectively implemented on the basis of the discipline proposed in [19]. If the flow is heterogeneous, the number of copies (replicas) of the request can be set depending on its limitations on the allowable expectation. Each copy of the request is sent to one of the cluster nodes. The generated query replicas are given different relative priorities. Nodes serving distributed replicas are selected randomly or cyclically. It is also possible that the same priorities are set for all waiting-critical requests.

It is of interest to study the methods of redundant maintenance when executing requests by a sequence of redundant nodes combined into a multi-level cluster. In [20], two variants of redundant maintenance of a heterogeneous flow are considered when copies of requests are sequentially passed through the nodes that make up the service paths. For the first option, a path is written for each copy as a sequence of nodes of different groups. The number of copies is set depending on the criticality of the request for service delays and does not change during the passage of the nodes that make up the path. For the second option, paths are formed dynamically at each stage, while a copy of the request executed first at some stage is transferred to the

next node group for redundant maintenance. Copies that are not served first at each stage are destroyed. At different stages of service, the redundancy multiplicity may vary. The assessment of the probability of the timeliness of sequential maintenance is carried out taking into account the accumulation of waiting in the nodes that make up the path of sequential execution of requests.

The functional heterogeneity of requests may lead to the need to complete cluster systems with nodes (servers) that are heterogeneous in parameters and functionality, combined into a cluster through communication nodes [21].

The analysis of fault-tolerant real-time cluster systems becomes more complicated when the input stream is heterogeneous in terms of the functionality of requests and their criticality to service delays. The analysis is also complicated by the functional heterogeneity of cluster nodes [21].

Reliability assessment taking into account the condition of timely servicing of requests in a cluster with functional heterogeneity of the flow and cluster nodes was considered in [21]. Cluster nodes according to [21] are functionally oriented servers that are connected through switching nodes. Requests through switching nodes are sent for service to the server nodes connected to them, corresponding functionality. At the same time, [21] does not take into account the delay in the switching nodes that distribute requests at the first stage of service. According to [21], the potential possibility of exceeding the maximum allowable total delay of two-stage maintenance already at the first stage of the distribution of the request through the switching node is also not taken into account. Software models [21] are not focused on clusters in which nodes combine a general-purpose server and functionally oriented servers. In such systems, the request is first served in a general-purpose server, and then sent for service to a functionally oriented server. Waiting in the queue of a general-purpose server can significantly affect the total delay in servicing requests and therefore they should be taken into account. You should also take into account the potential for exceeding the total waiting time already at the stage of servicing the request in a general-purpose server. The purpose of this work is to investigate the possibility of increasing the probability of timely execution in a cluster of a request flow that is heterogeneous in functionality and acceptable waiting time, taking into account the accumulation (summation) of waiting delays in queues at the service stages in a general-purpose server and in functionally oriented servers.

2. Cluster with functional heterogeneity of nodes

Consider a cluster combining m nodes. Each node of the cluster includes a general-purpose server (base server), with the possible connection of functionally oriented servers (functional modules) to it. The cluster structure is shown in Fig. 1.

Each incoming request is sent to one of the m cluster nodes (server groups). At the first stage, the request is served by a general-purpose server, and at the second stage by one of the functionally oriented servers of the cluster node. The system allocates n types of functional resources f_1, f_2, \dots, f_n . Cluster nodes can be equipped with different sets of functionally oriented servers, which may not coincide [21].

The functional capabilities of the cluster [19] are characterized by a matrix $\|\phi_{ij}\|_{m \times n}$, the element of which $\phi_{ij} = 1$, if the j -th node is equipped with workable resources focused on the execution of the f_i request, otherwise $\phi_{ij} = 0$, $j = 1, 2, \dots, m$; $i = 1, 2, \dots, n$.

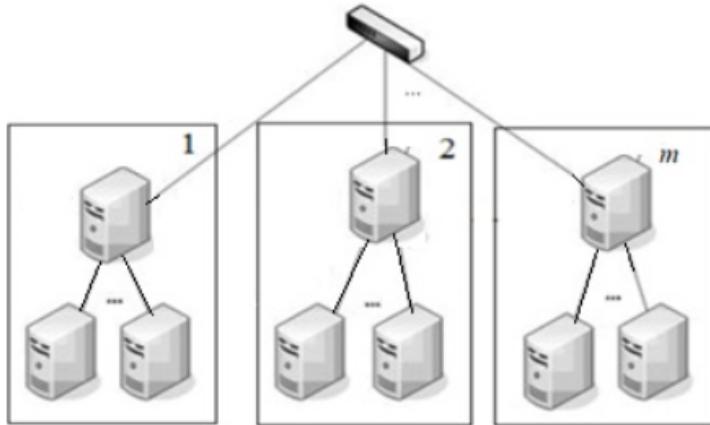


Fig. 1. Structure of a two-level cluster

3. Assessment of the timeliness of execution of requests of a functionally heterogeneous stream

In the cluster under consideration, the flow of requests of various functionality is first executed in a general-purpose server, and then in one of the functionally-oriented servers associated with it. Each server of a cluster node is represented by a queuing system of the M/M/1 type [22,23]. This representation of the node allows, based on the simplest models, to analyze the effectiveness of redundant servicing of requests of different functionality in two-tier clusters (clusters with two-stage servicing of requests by servers of two levels of the cluster). With a balanced distribution of the request flow between m base servers, we calculate the average waiting time for requests w as:

$$w = \left(\sum_{i=1}^n b_i \Lambda v_{0i}^2 / m \right) / \left(1 - \sum_{i=1}^n b_i \Lambda v_{0i} / m \right),$$

$$\sum_{i=1}^n b_i = 1,$$

where b_i and v_{0i} , respectively, are the share and average service time of the i -th functionality request in the base server of the cluster node, Λ is the intensity of the total request flow.

The probability of not exceeding the allowable waiting time t_i for i -functionality requests at the first stage of service in a general-purpose server is estimated based on a modification of the well-known formula [22] of the waiting time distribution of single-channel queuing systems $M/M/1$ with an infinite queue.

$$r_i = \left(1 - \frac{\Lambda v_0}{m} e^{(\frac{\Lambda}{m} - \frac{1}{v_0})t_i}\right),$$

where v_0 is the average query execution time in general purpose servers

$$v_0 = \sum_{i=1}^n b_i v_{0i}.$$

The modification takes into account the heterogeneity of the flow in terms of the allowable waiting time and the uniform distribution of requests across functional general-purpose servers. The probability of timely servicing of all types of functional requests f_1, f_2, \dots, f_n at two stages in general-purpose servers and functionally oriented servers is calculated as:

$$R = \prod_{i=1}^n \left(1 - \frac{\Lambda b_i v_i}{m_i} e^{(\frac{\Lambda b_i}{m_i} - \frac{1}{v_i})(t_i - w - v_{0i})}\right),$$

where v_i is the average execution time of the f_i request in the server of the i -th functionality.

The number of functionally oriented servers capable of executing the i -th functionality request is calculated as

$$m_i = \sum_{j=1}^m \phi_{ij}.$$

The performance indicator R corresponds to the requirement that requests of all n types of functionality must be completed in a timely manner. The indicator does not reflect the probability of requests of different functionality.

The average probability of timely two-stage maintenance of functional requests f_1, f_2, \dots, f_n taking into account the probability of their receipt, is calculated as:

$$A = \sum_{i=1}^n b_i r_i \left(1 - \frac{\Lambda b_i v_i}{m_i} e^{(\frac{\Lambda b_i}{m_i} - \frac{1}{v_i})(t_i - w - v_{0i})}\right).$$

4. Assessment of the timeliness of execution of requests of a functionally heterogeneous stream

When replicating requests of the i -th functionality with multiplicity k_i , we have

$$w = \left(\sum_{i=1}^n b_i k_i \Lambda v_{0i}^2 / m\right) / \left(1 - \sum_{i=1}^n b_i k_i \Lambda v_{0i} / m\right).$$

The probability of timely execution in a time less than t_i of at least one replica of the i -th functionality request

$$r_i = \left(1 - \frac{\Lambda \sum_{i=1}^n k_i b_i v_{0i}}{m} e^{(\frac{\Lambda \sum_{i=1}^n k_i b_i}{m} - \frac{1}{k_i b_i v_{0i}}) t_i}\right).$$

The probability of timely servicing of functional i_q functionality requests at two stages in general purpose servers and functionally oriented servers is calculated as:

$$R_i = \left\{1 - [1 - r_i \left(1 - \frac{k_i \Lambda b_i v_i}{m_i} e^{(\frac{k_i \Lambda b_i}{m_i} - \frac{1}{v_i})(t_i - w - v_{0i})}\right)]^k\right\}.$$

The probability of timely execution of all types of functional queries f_1, f_2, \dots, f_n with the multiplicity of their replication k_1, k_2, \dots, k_n is defined as

$$R = \prod_{i=1}^n R_i.$$

Average probability of timely execution of all types of requests

$$A = \sum_{i=1}^n b_i r_i.$$

5. An example of evaluating the timeliness of query execution when reserving the most delay-critical queries

Let's consider a variant of a fully functional configuration of cluster nodes when allocating two functional threads with an acceptable total query time in the system t_1 and t_2 . The reservation is carried out with a multiplicity of k only more critical to the expectation of requests.

The calculation will be carried out with the same average execution time of all types of requests v_0 . The average service delay in a general-purpose server with replication of requests more critical to waiting is calculated as

$$T_1 = v_0 / (1 - \Lambda(1 + b(k - 1))v_0/m),$$

The probabilities of timely execution of a replica (request) in the server for requests of the first and second type with an acceptable waiting time t_1 and t_2 are calculated as

$$r_1 = (1 - \frac{\Lambda(1 + b(k - 1))v_0}{m} e^{(\frac{\Lambda(1+b(k-1))}{m} - \frac{1}{v_0})t_1})$$

$$r_2 = (1 - \frac{\Lambda(1 + b(k - 1))v_0}{m} e^{(\frac{\Lambda(1+b(k-1))}{m} - \frac{1}{v_0})t_2})$$

The probability of timely servicing of a heterogeneous flow with two gradations of the criticality of requests to an acceptable expectation, taking into account the obligation to timely fulfill all types of requests for functionality, is estimated as

$$R = (1 - \{1 - r_1(1 - [[k\Lambda bv_1/m]e^{(\frac{k\Lambda b}{m} - \frac{1}{v_1})(t_1 - T_1)}])\}^k) \times \\ \times \{r_2(1 - [(\Lambda(1 - b)v_2)/m]e^{(\frac{\Lambda(1-b)}{m} - \frac{1}{v_2})(t_2 - T_1)})\}.$$

The average probability of timely execution of requests of various functionality is calculated as

$$r_2 = (1 - \frac{\Lambda(1 + b(k - 1))v_0}{m} e^{(\frac{\Lambda(1+b(k-1))}{m} - \frac{1}{v_0})t_2})$$

$$A = b\{1 - \{1 - r_1(1 - [[k\Lambda bv_1/m]e^{(\frac{k\Lambda b}{m} - \frac{1}{v_1})(t_1 - T_1)}])\}^k\} + \\ + (1 - b)\{r_2(1 - [(\Lambda(1 - b)v_2)/m]e^{(\frac{\Lambda(1-b)}{m} - \frac{1}{v_2})(t_2 - T_1)})\}.$$

The calculation will be carried out at $b = 0.5$ and $t_2 = 1$ s, $m = 5$ nodes, the average service time of any request in the general server $v_0 = 0.01$ s, and in the functional server $v_1 = v_2 = v = 0.1$ s.

The dependence of the probability of timely service on the intensity of a heterogeneous flow of requests according to criterion R and A when duplicating more delay-critical requests is shown in Fig. 2 a and b.

Curves 1-3 correspond to $t_1 = 0.1; 0.15; 0.2$ s when duplicating more critical queries, and curves 4-6 without reserving them. The dependence of the desired probabilities R and A on the multiplicity of reserving more critical to waiting

requests are shown in Fig. 3 a and b. Curves 1-3 correspond to the input flow intensities $\Lambda = 10, 15, 20 \text{ 1/s}$ at a fraction of $b = 0.5$, and curves 4-6 at $b = 0.7$.

The presented dependencies show the existence of an area of efficiency of reservation of latency-critical requests and the existence of an optimal multiplicity of their reservation of critical requests.

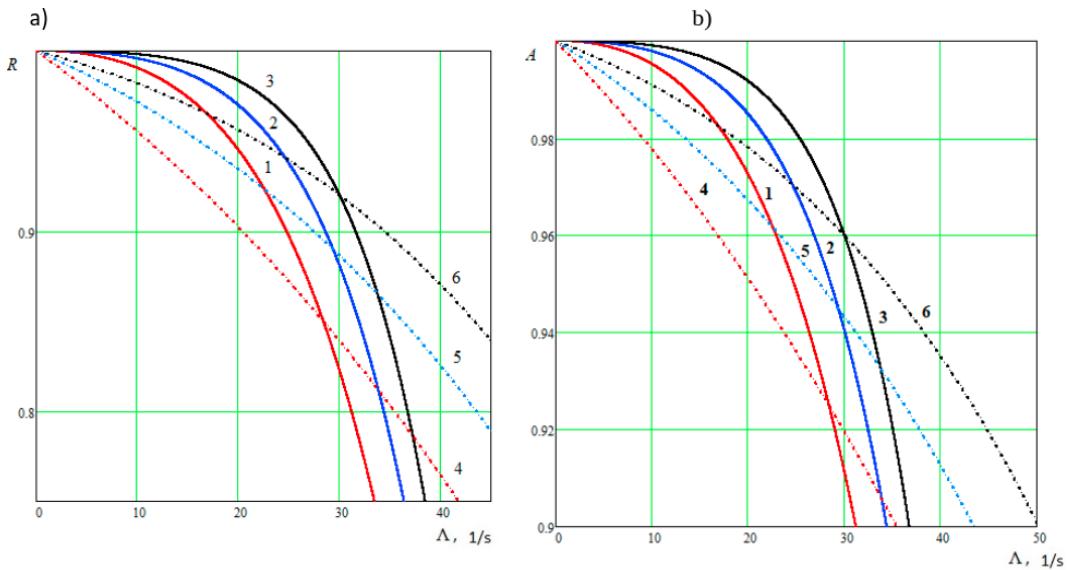


Fig. 2. Dependence of the probabilities of timely maintenance on the intensity of R (a) and A (b)

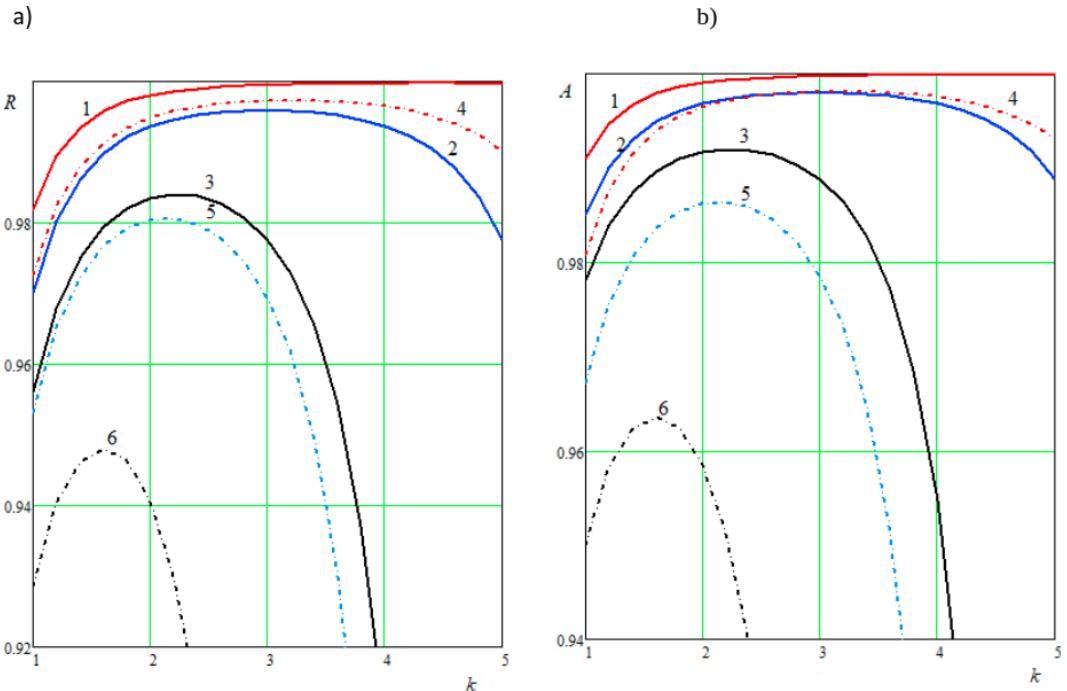


Fig. 3. Dependences of the probabilities of timely service R (a) and A (b) on the multiplicity of reservations of more critical to waiting requests

6. Conclusion

For clusters with functional heterogeneity of nodes, an analytical model is proposed for estimating the probability of timely execution of requests of a flow that is heterogeneous in functionality and acceptable waiting time. The timeliness of query execution is determined by taking into account the accumulation of waiting delays in queues during two-stage maintenance, first in a general-purpose server, and then in one of the functionally oriented servers.

For the two-level clusters under consideration, the possibility of increasing the probability of timely execution of requests of a stream that is heterogeneous in functionality and acceptable waiting time is shown.

The existence of an optimal redundancy multiplicity of the most delay-critical functional requests is shown.

The area of efficiency of redundant service of requests critical to delays is determined depending on the maximum allowable waiting time in nodes sequentially executing the request, on the intensity of the total input stream and on the shares of streams of various functionality.

REFERENCES

1. Aab A. V., Galushin P. V., Popova A. V., Terskov V. A. Mathematical model of reliability of information processing computer appliances for real-time control systems. Siberian Journal of Science and Technology. 2020, Vol. 21, No. 3, P. 296–302. Doi: 10.31772/2587-6066-2020-21-3-296-302
2. Sorin D. Fault Tolerant Computer Architecture. Morgan Claypool 2009. 103 p.
3. Pavsky V.A., Pavsky K.V. Mathematical model with three-parameters for calculating reliability indices of scalable computer systems . 2020 International Multi-Conference on Industrial Engineering and Modern Technologies, FarEast-Con 2020. 2020. p. 9271332
4. Malik V., Barde C.R. Live migration of virtual machines in cloud environment using prediction of CPU usage // International Journal of Computer Applications. 2015. V. 117 N 23. P. 1–5. doi: 10.5120/20691-3604
5. Houankpo H.G., Kozyrev D.V, Nibasumb E. M., Mouale N.B. Mathematical Model for Reliability Analysis of a Heterogeneous Redundant Data Transmission System International Congress on Ultra Modern Telecommunications and Control Systems and Workshops 2020-October,9222431, . 189-194.
6. Tatarnikova T.M., Poyanova E.D. Differentiated capacity extension method for system of data storage with multilevel structure. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2020, 20(1), . 66–73 DOI: 10.17586/2226-1494-2020-20-1-66-73
7. Astakhova T. N., Verzun N. A., Kasatkin V. V. , Kolbanov M. O., Shamin A. A. Sensor network connectivity models. Informatsionno-upravliaiushchie sistemy [Information and Control Systems], 2019, no. 5, pp. 38–50. doi:10.31799/16848853-2019-5-38-50.
8. Zakoldaev D.A., Korobeynikov A.G., Shukalov A.V., Zharinov I.O. Workstations Industry 4.0 for instrument manufacturing //IOP Conf. Series: Materials Science and Engineering665 (2019) 012015IOP Publishingdoi:10.1088/1757-899X/665/1/012015
9. Kutuzov O. I., Tatarnikova T. M. Model of a self-similar traffic generator and evaluation of buffer storage for classical and fractal queuing system. In Moscow Workshop on Electronic and Networking Technologies, MWENT 2018 - Proceedings 1, pp. 1-3 (2018).
10. Jiang Z., Wang G., He Y., Huang Z., Xue R. Reliability prediction for computer numerical control machine servo systems based on an ipso-based rbf neural network Shock and Vibration. 2022. v. 2022. p. 2684942
11. Gómez C.E., Chavarriaga J., Castro H.E. Reliability analysis in desktop cloud systems .Communications in Computer and Information Science. 2020. v. 1277 CCIS. p. 165-180

12. Mukhin V., Loutskii H., Kornaga Y., Steshyn V., Barabash O. Models for analysis and prognostication of the indicators of the distributed computer systems' characteristics. International Review on Computers and Software. 2015. v. 10. № 12. p. 1216-1224.
13. Kim I S., Choi Y., Constraint-aware VM placement in heterogeneous computing clusters. //Cluster Comput. 23, 71–85 (2020). <https://doi.org/10.1007/s10586-019-02966-6>.
14. Newman M., de Castro L.A., Brown K.R. Generating fault-tolerant cluster states from crystal structures. Quantum. 2020. v. 4
15. Tourouta E., Gorodnichev M., Polyantseva K., Moseva M. Providing fault tolerance of cluster computing systems based on fault-tolerant dynamic computation planning. Lecture Notes in Information Systems and Organisation. 3rd. "Digitalization of Society, Economics and Management - A Digital Strategy Based on Post-pandemic Developments" 2022. p. 143-150
16. Bukhsh M., Abdullah S., Arshad H., Rahman A., Asghar M.N., Alabdulatif A. An energy-aware, highly available, and fault-tolerant method for reliable IoT systems. IEEE Access. 2021. v. 9. p. 145363-14538
17. Shcherba E.V., Litvinov G.A., Shcherba M.V. Securing the multipath extension of the olsr routing protocol: 13th International IEEE Scientific and Technical Conference Dynamics of Systems, Mechanisms and Machines, Dynamics 2019 - Proceedings. 2019. . 8944743
18. Lee M.H., Dudin A.N., Klimenok V.I. The SM/V/N queueing system with broadcasting service // Math. Probl. in Engineer. 2006. V. 2006. Article ID 98171. 18 p.
19. Bogatyrev S.V., Bogatyrev A.V., Bogatyrev V.A. Priority Maintenance with Replication of Wait-Critical Requests. Wave Electronics and its Application in Information and Telecommunication Systems (WECONF 2021). 2021. pp. 9470640.
20. Bogatyrev V.A. , Bogatyrev A.V. , Bogatyrev, S.V. : Redundant Servicing of a Flow of Heterogeneous Requests Critical to the Total Waiting Time During the Multi-path Passage of a Sequence of Info-Communication Nodes. Lecture Notes in Computer Science// Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020. Vol. 12563. pp. 100-112. DOI 10.1007/978-3-030-66471-8_9
21. Bogatyrev V.A. , Bogatyrev A.V., Bogatyrev, S.V, Reliability and probability of timely servicing in a cluster of heterogeneous flow of query functionality // Wave Electronics and its Application in Information and Telecommunication Systems (WECONF 2020) - 2020, pp. 9131165 DOI 10.1109/WECONF48837.2020.9131165

22. Kleinrock L. Queueing Systems: Volume I. Theory. New York: Wiley Inter-science.1975 p. 417. ISBN 978-0471491101
23. Vishnevsky V.M. Theoretical foundations of computer network design Author: Publisher: Technosphere ISBN: 5-94836-011-3 : 2003

UDC: 004.67

Neuro-Fuzzy Model Based on Multidimensional Membership Function

T.A. Bui¹, F.F. Pashchenko², D.H. Tran¹, V.T. Nguyen¹, T.N. Pham³

¹Moscow Institute of Physics and Technology (State University), Institutskiy Pereulok, 9, Dolgoprudny, Moscow Oblast, Russia

²Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Ulitsa Profsoyuznaya 65, Moscow, Russia

³Ho Chi Minh University of Natural Resources and Environment, 236B Le Van Sy, Ward 1, Tan Binh District, Ho Chi Minh City, Vietnam

buitruonganmta93@gmail.com, pif-70@yandex.ru, imroytran@phystech.edu,
vnt.702@gmail.com, phamnguyenvnuf2014@gmail.com

Abstract

Currently, systems based on multidimensional membership functions are being actively researched and developed. Most algorithms for determining the parameters of fuzzy membership functions are developed on the basis of one-dimensional membership functions. The fuzzy rules generated by these algorithms often overlap and cannot act as independent rules. The overlap of fuzzy rules in fuzzy systems does not allow one to evaluate the reliability of individual fuzzy rules and at the same time creates limitations in extracting knowledge from fuzzy systems. In this article, a neuro-fuzzy neural system will be built based on a multidimensional Gaussian membership function with the ability to describe the relationship of interaction between input variables, and at the same time, the generated fuzzy rules are capable of independent operation.

Keywords: fuzzy set, multidimensional membership functions, Gaussian functions, fuzzy inference system, neuro-fuzzy model, Kaczmarz algorithm, recurrent least squares.

1. Introduction

Fuzzy logic is often used in conjunction with neural networks to create systems that have the benefits of fuzzy logic clarity and neural network learning capabilities. Such neuro-fuzzy systems are very diverse and are increasingly being improved in accordance with the development of neural network learning algorithms.

Most of the above works are built on the basis of one-dimensional membership functions, such as Gaussian, Bell, Triangular, RBF. The limitation of this approach

is the complexity of the model in terms of the number of rules, which increases exponentially with the number of inputs (spatial curse).

As an effective solution to the above problem, the use of multidimensional membership functions in a fuzzy inference system is proposed. The most commonly used multivariate membership function is the multivariate Gaussian function [1-3]. Gaussian functions require fewer parameters to describe a fuzzy rule than linear membership functions [4].

The above models of multidimensional fuzzy neural networks show the relationship between input variables. However, the generated fuzzy rules often overlap. This affects the independent operation of each fuzzy rule, which leads to inefficient data extraction from the fuzzy model.

2. Neuro-fuzzy model based on multidimensional membership function

The multidimensional Gaussian membership function (Fig. 1) can be described as [5]:

$$H(x_0) = e^{-(x_0 - C)S^{-1}(x_0 - C)^T}, \quad (1)$$

where C is the center of the membership function, S is the matrix of expansion coefficients. The matrix S is initialized and folded as a positive-definite matrix, which ensures that the value of $H(x_0)$ is always within the half-interval (0,1].

The proposed architecture of the model consists of 4 layers shown in Fig. 2.

Layer 1. Calculation of the degree of membership of the input vector for each fuzzy rule

$$H_i = e^{-(X_t - C_i)S_i^{-1}(X_t - C_i)^T}, i = \overline{1, m}, \quad (2)$$

where m is the number of fuzzy rules.

Layer 2. Finding the normalized rule truth value

$$\beta_i = \frac{H_i}{\sum_{k=1}^m H_k}, i = \overline{1, m}, \quad (3)$$

Layer 3. Calculation of the parameters of the consequences $\beta_i Y_{i,t}$

Depending on the type of model, the value of $Y_{i,t}$ can be a constant

$$Y_{i,t} = c_i, i = \overline{1, m}, \quad (4)$$

or a linear function

$$Y_{i,t} = c_{i,0} + c_{i,1}X_{t,1} + c_{i,2}X_{t,2} + \dots + c_{i,n}X_{t,n}, i = \overline{1, m}, \quad (5)$$

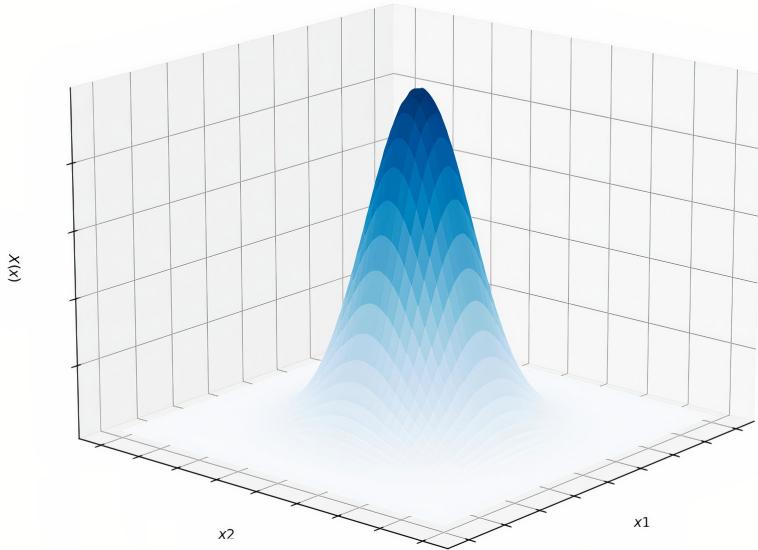


Fig. 1. Gaussian membership function with two variables

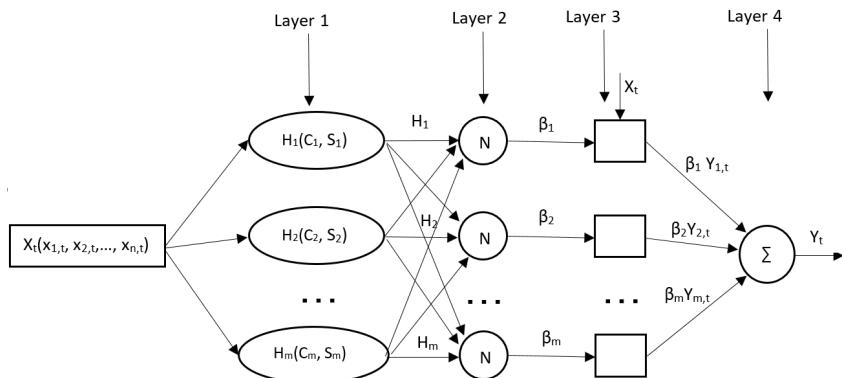


Fig. 2. Architecture of the neuro-fuzzy model based on the MMF

where n is the dimension of the input space.

Layer 4. The sum of all incoming values from the 3rd layer

$$Y_t = \sum_{i=1}^m \beta_i Y_{i,t}, \quad (6)$$

Fuzzy neural model synthesis algorithm:

The idea of the fuzzy neural model synthesis algorithm is to divide the input space with each direction divided into n equal segments, and then consider pairs of fuzzy rules with distances less than a given value dmin.

If two fuzzy rules are spaced less than dmin and do not give a rapid increase in the error function, then they are pooled together. The algorithm is stopped when the model achieves a sufficiently small number of fuzzy rules or the model error exceeds the threshold.

3. Approach to the application of algorithms for identifying the coefficients of linear equations

The output value of the neuro-fuzzy model can be described as follows:

$$\hat{y}(t) = \sum_{i=1}^m \beta_i(t) Y_i(t) = \sum_{i=1}^m \sum_{j=0}^n c_{i,j} \cdot \beta_i(t) \cdot X_j(t), \quad (7)$$

where n is the dimension of the input space, $c_{i,j}$ - is the j-th free coefficient of the i-th fuzzy rule, $X_j(t)$ is the value of the j-th input at the t-th observation ($X_0(t) = 1$ for any observation t).

Let $= (c_{1,0}, \dots, c_{m,0}, c_{1,1}, \dots, c_{m,1}, \dots, c_{1,n}, \dots, c_{m,n})^T$ - vector of adjustable coefficients.

$\tilde{x}(t) = (\beta_1(t), \dots, \beta_m(t), \beta_1(t) \cdot X_1(t), \dots, \beta_m(t) \cdot X_1(t), \dots, \beta_1(t) \cdot X_n(t), \dots, \beta_m(t) \cdot X_n(t))^T$ - the value of the extended input vector at observation t.

Then (1) can be written as follows:

$$\hat{y}(t) = \tilde{c}\tilde{x}(t), \quad (8)$$

In this case, the task of identification is reduced to finding the parameters minimizing the objective functional:

$$J(c) = M \left\{ L \left(\hat{X}, Y, \hat{Y}, c \right) \right\}, \quad (9)$$

where M is the expectation symbol, L is the loss function, \hat{X} is the augmented input vector, Y is the target vector, and \hat{Y} is the output vector of the model. The desired value c is the solution to the following problem:

$$c^* = \arg \min_c J(c), \quad (10)$$

In the following, we will explore commonly used methods for finding the value of c^* .

3.1. Recurrent least squares. According to this method, the value of c is determined by the following recursive formula [6]:

$$P(t) = P(t-1) - \frac{P(t-1)\tilde{x}(t)\tilde{x}^T(t)P(t-1)}{1 + \tilde{x}^T(t)P(t-1)\tilde{x}(t)} \quad (11)$$

$$c(t) = c(t-1) + P(t-1)(y(t) - c^T(t-1)\tilde{x}(t)) \quad (12)$$

where $P(t)$ is a correction matrix of size $m(n+1) \times m(n+1)$,

$$P(0) = \begin{bmatrix} \alpha & 0 & \dots & 0 \\ 0 & \alpha & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \alpha \end{bmatrix} \quad (13)$$

α is a large enough number.

The value to be determined is $c* = c(N)$, where N is the number of observations.

3.2. Kaczmarz algorithm. According to the Kaczmarz algorithm, c is determined by the formula [6]:

$$c(t) = c(t-1) + \frac{y(t) - \hat{y}(t)}{\|\tilde{x}(t)\|^2}\tilde{x}(t). \quad (14)$$

3.3. Generalized adaptive identification algorithms. The general form of this algorithm can be represented by the formula [6]:

$$c(t) = \alpha(t)c(t-1) + \beta(t)\tilde{x}(t) \quad (15)$$

where:

$$\alpha(t) = \frac{\|c(t-1)\|^2\|\tilde{x}(t)\|^2 - y(t)(c^T(t-1)\tilde{x}(t))}{\|c(t-1)\|^2\|\tilde{x}(t)\|^2 - (c^T(t-1)\tilde{x}(t))^2} \quad (16)$$

$$\beta(t) = \frac{\|c(t-1)\|^2(y(t) - c^T(t-1)\tilde{x}(t))}{\|c(t-1)\|^2\|\tilde{x}(t)\|^2 - (c^T(t-1), \tilde{x}(t))^2}$$

To compare the rate of convergence and performance of the algorithms, we use two examples of identifying the coefficients of systems with the following parameters:

$X = [x_0, x_1, x_2, x_3]$ - vector of input data

$$x_0(i) = 1$$

$$x_1(i) = 1 + 0.4 \sin(0.1i\pi + 0.2)$$

$$x_2(i) = 1.5 + \sin(0.05i\pi + 0.4)$$

$$x_3(i) = 1 + \sin(0.07i\pi + 1.6)$$

where $i = 1..300$.

$K = [k_1, k_2, k_3, k_4]$ - vector of object parameters.

$k_1 = 2.5$ for a stationary system and $k_1 = 2 + 0.5 \sin(0.001i\pi + 0.2)$ for a non-stationary system

$$k_2 = -1.7; k_3 = 1.3; k_4 = 0.25$$

The simulation results of the above methods for identifying coefficients for a stationary model are shown in Fig. 3.

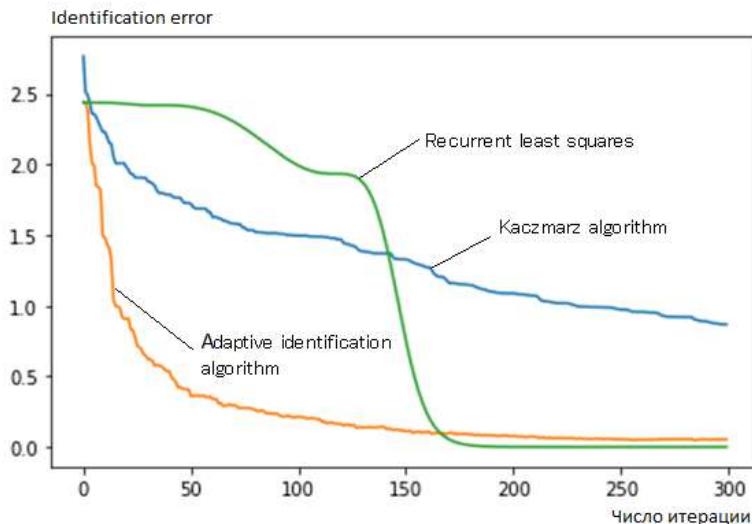


Fig. 3. Uncertainties in Estimating the Parameters of a Stationary System

The running time of the algorithms, respectively:

Recurrent least squares: 0.26s

Generalized adaptive identification algorithms: 0.19

Kaczmarz algorithm: 0.13

On fig. 4 shows the results of modeling identification methods for a non-stationary system.

The running time of the algorithms, respectively:

Recurrent least squares: 0.28s

Generalized adaptive identification algorithms: 0.21

Kaczmarz algorithm: 0.13

The simulation results show that the Kaczmarz algorithm has the shortest execution time. However, the disadvantage of this method is the rate of convergence. The least squares method provides a high rate of convergence, but has the disadvantage

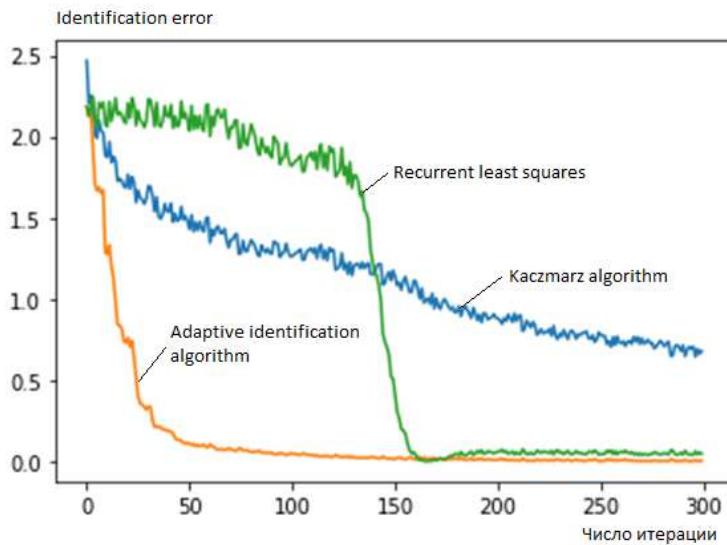


Fig. 4. Uncertainties in Estimating the Parameters of a Nonstationary System

of computational complexity, which leads to a long implementation time. The generalized adaptive identification algorithm has a shorter execution time than the least squares algorithm and has a high convergence rate.

Based on the results of the analysis, an approach to the use of algorithms for identifying the coefficients of linear equations is proposed:

For an available data set, a neuro-fuzzy model does not require a high learning rate. In this case, using the recurrent least squares method allows you to achieve optimal accuracy.

For data collected sequentially, the model needs to be updated in a timely manner in accordance with new changes in the input data. The coefficient homogenization algorithm must simultaneously meet the requirements of processing speed and accuracy. In this case, a generalized adaptive identification algorithm should be used.

4. Building a fuzzy neural model to predict the exchange rate

In this section, we apply a fuzzy neural model based on a multivariate membership function to solve the exchange rate forecasting problem.

To build a model, it is necessary to collect data on the factors that affect the exchange rate. Then use the information variable selection algorithm to select the best factors for the model.

The following factors influence the exchange rate:

- oil price
- gold price
- import
- export

The input data is the value of the exchange rate and 5 factors that affect it for 62 consecutive months from January 2014 to February 2019 [7].

For complex models, using the maximum correlation coefficient instead of building the model at each step will greatly reduce the computational cost of attribute selection. The search algorithm for information variables will look like this [8]:

- Select the original model M_0 .
- Calculate the maximum correlation coefficient between input variable X and output variable Y.
- Add the variable most correlated with the output variable to the M_0 model to get the M_1 model.
- We build model M_1 with the selected input variable.
- Based on the maximum correlation coefficient between the current model error M_1 and the rest of the input variables, select the next variable.
- Add the selected variable to M_1 to get M_2 .
- Let the model M_k be built. We select the next input variable based on the maximum correlation coefficient between the remaining input variables and the model error M_k . The selected variable is added to M_{k+1} .
- Repeat until all required input variables are included in the model.
- The algorithm stops if it finds a model M_k that satisfies the given stopping condition. Otherwise, the loop continues until all required input variables in the model have been combined.
- We select from the models $\{M_1, \dots, M_k, \dots, M_n\}$ the model that provides the best accuracy for the selected quality metric.

The information variable selection algorithm stops at two inputs: the price of gold and the price of oil. The model is initialized with the following parameters:

$$n_{term} = 30$$

$$loss_{threshold} = 0.001$$

$$loss_{max} = 0.5$$

The error function used is LSE:

$$LSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (17)$$

The right side of the model is a linear equation

$$Y_{i,t} = c_{i,0} + c_{i,1}X_{t,1} + c_{i,2}X_{t,2} + \dots + c_{i,n}X_{t,n}, i = \overline{1, m}, \quad (18)$$

The number of fuzzy rules after training is 15.

The model training error is 0.601.

The model testing error is 0.842.

Model error after additional learning: 0.81.

The model has been built in Python programming language on the Google Colab Free.

5. Conclusion

This paper presents a method for constructing fuzzy neural models based on multidimensional membership functions. A method for selecting input variables and an approach to the application of algorithms for identifying the coefficients of linear equations have been developed. Algorithms for solving the problem of building a model for forecasting the exchange rate are applied. The resulting model is able to describe the interaction between input variables, and the generated fuzzy rules are able to work independently. This is an important prerequisite for extracting knowledge from digital data.

REFERENCES

1. Dongyeop Kang, Woojong Yoo, Sangchul Won “Multivariable TS fuzzy model identification based on mixture of Gaussians” 2007 International Conference on Control, Automation and Systems, December 2007.
2. Andre Lemos, Walmir Caminhas, Fernando Gomide “Multivariable Gaussian Evolving Fuzzy Modeling System” IEEE Transactions on Fuzzy Systems Vol. 19, pp. 91 – 104, October 2010.
3. Mahardhika Pratama, Sreenatha. G. Anavatti, Plamen. P. Angelov, Edwin Lughofer “PANFIS: A Novel Incremental Learning Machine” IEEE Transactions on Neural Networks and Learning Systems, Vol. 25, pp. 55 – 68, July 2013.
3. Mahardhika Pratama, Sreenatha. G. Anavatti, Plamen. P. Angelov, Edwin Lughofer “PANFIS: A Novel Incremental Learning Machine” IEEE Transactions on Neural Networks and Learning Systems, Vol. 25, pp. 55 – 68, July 2013.
4. J. Abonyi, R. Babuska, F. Szeifert “Fuzzy modeling with multivariate membership functions: gray-box identification and control design” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) Vol. 31, pp. 755 – 767, Oct. 2001.
5. Mahardhika Pratama, Sreenatha. G. Anavatti, Plamen. P. Angelov, Edwin Lughofer “PANFIS: A Novel Incremental Learning Machine” IEEE Transactions on Neural Networks and Learning Systems, Vol. 25, pp. 55 – 68, July 2013.

6. Galina Pikina, Alexander Pashchenko, Artur Avetisyan, Gennady Filippov, F. F. Pashchenko. Thermal-hydraulic models of power station equipment. M.-FIZMATLIT, 2013. 448 p.
7. Bui Tr.A., Pashchenko F. F., Pashchenko A. F., Kudinov Y.I. Using Neural Networks To Predict Exchange Rates // 2020 13th International Conference "Management of large-scale system development" (MLSD), 2020. – P. 1-5.
8. Kamenev Andrey V., Pashchenko Alexandr F., Pashchenko Fedor F. Neuro-fuzzy simulation system with the selection of informative variables//SENSORS SYSTEMS. No 7-8. -P. – 8-14.

UDC: 004.94

Performance analysis of a finite-source retrial queueing system with two-way communication, catastrophic breakdown and impatient customers using simulation

János Sztrik¹, Ádám Tóth¹, Ákos Pintér¹, Zoltán Bács¹

¹University of Debrecen, Debrecen 4032, Hungary

{toth.adam,sztrik.janos}@inf.unideb.hu, bacs.zoltan@econ.unideb.hu,
apinter@science.unideb.hu

Abstract

An M/M/1//N retrial queueing system with two-way communication to the infinite source and impatient customers in the orbit is considered in the paper. There is a finite source in which the primary or regular customers are coming, while requests from the infinite source are the secondary customers. When the service unit is in an idle state, it may call a customer from the infinite source for service. Besides, the server is supposed to break down according to several distributions which have a specialty in removing all the customers located in the system. In the case of a faulty state, blocking is applied not allowing the customers into the system until the service unit fully recovers. This work concentrates on examining the effect of those distributions on several performance measures. The obtained results are graphically realized to show the differences and curiosities among the used parameter settings of the various distributions.

Keywords: Simulation, Catastrophic breakdown, Retrial queuing system, Collision, Impatience, Sensitivity analysis

1. Introduction

In many fields of our life, the phenomenon of waiting appears making inevitably to create queueing systems handling for example increasing network traffic in many info-communication systems. Throughout the years researchers have designed numerous tools and mechanisms which are suitable for modeling various organizations and one prime instance is retrial queueing systems that are capable of depicting arising real-life problems in telecommunication schemes like telephone switching systems, call centers, computer networks, and computer systems. Several publications exist

where researchers exploit the advantages of retrial-queueing systems with repeated calls using for their models like in [1].

In the case of a retrial queueing system, a virtual waiting room is taken into consideration which is called the orbit meaning that whenever a service of a job can not start because of failure or occupation of the server it remains in the system. In the orbit, these customers have the opportunity to be at the service facility after a random time. The population of the customers is finite as the probability that a server calls a customer from the orbit is not very small and under such circumstances, it is more suitable to examine models with retrial queues. Exploring the available literature many paper have applied infinite and finite source queueing systems for example in [2].

It is also interesting to observe how the feature of two-way communication is used in papers in many fields of life. Its popularity originates from its usefulness to model applicable systems creating real-life applications. One prime example can be mentioned in the topic of telecommunication, especially in call centres where agents may be occupied with other particular labor during in an inactivity period like selling, advertising, and promoting products besides handling the calls of the customers. Optimizing the utilization of the service units or agents is always pivotal to increasing the efficiency of such systems. To mention some works about applying two-way communication schemes here are some instances [3], [4].

Random breakdowns and malfunctions occur in real-life scenarios caused by a power outage, human negligence, or other catastrophic activity. This greatly changes the operation of the system and the performance measures thus its investigation is necessary. In many cases, the service units are presumed to be accessible all the time which is not realistic. Systems with random failures have been investigated by many researchers for example in [5] [6]. However, there are certain situations where the effect of breakdown is different types of failures can be investigated. For instance, power outages or mechanical failures may cause catastrophic events in which all the customers in the system are removed. This is known as a negative customer and it takes out every other request from service upon its arrival. This eventuates a disaster event because it also breaks down the service unit and in this case, every customer is forwarded back to the source. Papers in connection with negative customers can be found in [7], [8].

The aim of this work is to realize a sensitivity analysis using various distributions of failure time on several performance measures while the departure of customers may happen. The results are obtained by our stochastic simulation program using the basics of SimPack ([9]) which contains the basic building blocks of a simulation model. This gives us the opportunity to model any type of queueing system to create any type of simulation model and we can calculate any performance measure

using arbitrary random number generators for the desired random variable. The presented curves highlight both the effect of disaster events and the impatience of the customers applying various parameter settings and these figures concentrate on the interesting phenomena of these systems. The table of input parameters and graphical illustrations of the results are included demonstrating the influence of the used distributions on the main performance metrics.

2. System model

A finite-source retrial queueing system of type $M/M/1//N$ is regarded with an unreliable service unit and impatient customers. This model has a service unit and exactly N individual resides in the finite-source in which request generation (primary request) is proceeded towards the system according to exponential law with parameter λ . This means that the inter-arrival times are exponentially distributed with mean λ . If an arriving job finds the server in an idle state then its service starts immediately which is an exponentially distributed variable with μ_1 . Otherwise, in vain of a queue, jobs are not lost but remain in the system being forwarded to the orbit which is a virtual waiting space. From there these requests after an exponentially random time with parameter ν retry to reach the service unit. After spending futilely an exponentially distributed time with rate τ in the orbit a customer may choose to leave the system without being served so in other words, every request has an impatience characteristic. It is assumed random breakdowns take place according to various distributions like gamma, hyper-exponential, Pareto, hypo-exponential, and lognormal. The parameters are chosen in a way that a real sensitivity analysis would be accomplished. In these occurrences, disaster events develop resulting in interruption of the service of a job and with the customers, in the orbit, they all depart from the system. Blocking is applied during faulty periods so no customers are allowed by the system until the server fully recovers.

The repair process begins to be executed after the failure of the service unit which happens according to an exponential distribution with parameter γ_2 . Two-way communication was also introduced in our model, when the server becomes free it may call a request from an infinite source (secondary customer) after an exponentially random time with parameter λ_2 . That type of customer occupies instantly the service unit if it is not busy upon its arrival, otherwise, it is forwarded to a special buffer where it waits there until the server turns idle. At that moment a secondary customer automatically enters the service facility. In the case of a catastrophic event, every primary job returns to the finite-source, and every secondary customer exits from the system including the one who is under service. The service time of the secondary customer also follows an exponential distribution with the rate of μ_2 . Every appeared

arbitrary variable in the model construction is supposed to be independent of each other.

3. Simulation results

As mentioned earlier SimPack was used as a base of our program and we used a statistical package that was responsible for obtaining the desired performance measures.

The confidence level of 99.9% is employed throughout the simulations and 0.00001 is the amount of the relative half-width of the confidence interval to pause the actual simulation sequence. The size of the batch in the initial transient period can not be too small therefore its value is set to 1000.

In Table 1 the used values of input parameters are presented.

Table 1. Numerical values of model parameters

N	ν	μ_1	μ_2	τ	γ_2	λ_2
100	0.01	1	2.5	0.05	1	0.5

The next table (Table 2) consists of the parameters of failure time, every chosen parameter is according to have the same mean and variance value in that way a valid comparison is achieved. The simulation program was tested by many parameter values and the reason for selecting these values is focusing on interesting situations besides that it is worth mentioning that almost the same phenomenon appeared as with this particular setting. The squared coefficient of variation is more than one in this scenario which is totally intentional to check the influence of peculiar random variables.

Table 2. Parameters of failure time

Distribution	Gamma	Hyper-exponential	Pareto	Lognormal
Parameters	$\alpha = 0.31225$ $\beta = 0.05588$	$p = 0.36197$ $\lambda_1 = 0.12955$ $\lambda_2 = 0.22835$	$\alpha = 2.1455$ $k = 2.9835$	$m = 1.00278$ $\sigma = 1.19819$
Mean	5.558			
Variance	100			
Squared coefficient of variation	3.2024857438			

Due to the page limitation we involved just two figures, the others and the results of other scenario will be displayed in the extended version of the paper. In Figure 1a i represents the number of primary customers in the system on the X-axis, and $P(i)$ denotes the probability that exactly i primary customers are situated at the server and in the orbit altogether on the Y-axis. The distribution of the number of primary customers in the system is displayed when λ is 0.11 using various distributions of failure time. The mean number of primary customers in the system differs from each

other greatly. In the case of the gamma distribution, customers tend to spend more time in the system compared to Pareto distribution. It is also noticeable that the highest probability is at 0 and this can be explained by the fact that during faulty periods customers are not allowed to enter and for every catastrophic breakdown the system is emptied.

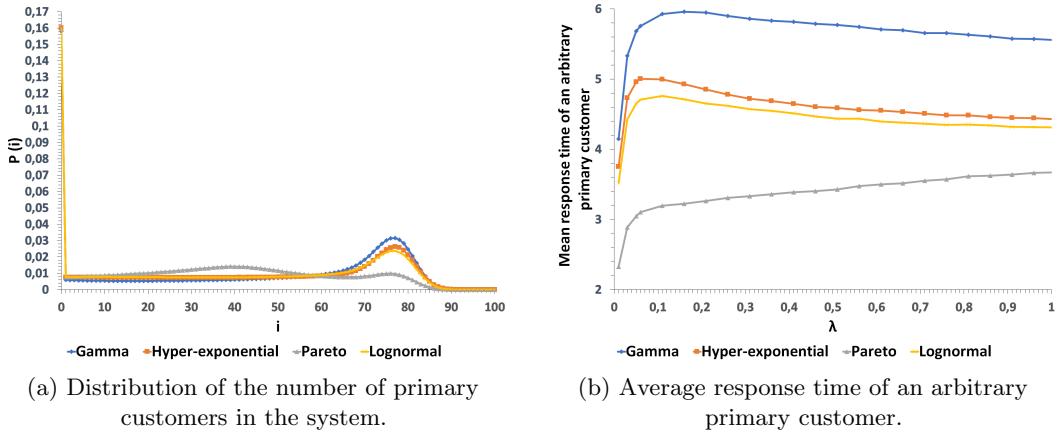


Fig. 1. The effect of different distributions

The expected response time of an arbitrary primary customer is presented in the function of the arrival intensity of incoming primary request in Figure 1b. Even though the mean and the variance value are equal with each other, huge gaps develop among the applied distributions, at gamma the highest average response times can be observed compared to the others. Also with the increment of the arrival intensity, the expected response time of an arbitrary primary customer starts to increase then after a certain intensity arrival ($\lambda = 0.07$) it decreases except in the case of Pareto.

Although the mean operation time is 5.558 the values of the expected response times are higher than that which is quite interesting. Our intuition is that this can be explained by that the variance is quite high resulting in many small operation times and in most of them no customer can enter because they are so small. But there are several high operation periods in which it is very probable that many jobs enter and spend relatively a high amount of time.

4. Conclusion

A finite-source retrial queueing system is introduced with a non-reliable server, a two-way communication scheme, and impatient customers. We investigated a scenario where different parameters are used to carry out a sensitivity analysis to

figure out how the different performance measures develop. The results are obtained by our simulation program and several graphical figures depict the effect of using various distributions of failure time on the expected response time of primary and on the distribution of customers in the system. In our figures, the differences were clearly seen among the values of several performance measures when the squared coefficient of variation is greater than one showing how pivotal applying a distribution. The curves also reveal the impact of impatience on reducing the average response time of a primary customer. In the future we plan to continue our research work, examining other types of finite-source retrial queuing systems with two-way communication or adding another service unit for backup purposes.

REFERENCES

1. J. Kim, B. Kim, A survey of retrial queueing systems, *Annals of Operations Research* 247 (1) (2016) 3–36.
2. T. Takeda, T. Yoshihiro, A distributed scheduling through queue-length exchange in CSMA-based wireless mesh networks, *Journal of Information Processing* 25 (2017) 174–181.
3. V. Dragieva, T. Phung-Duc, Two-way communication $M/M/1/N$ retrial queue, in: International Conference on Analytical and Stochastic Modeling Techniques and Applications, Springer, 2017, pp. 81–94.
4. J. Sztrik, Á. Tóth, Á. Pintér, Z. Bács, The simulation of finite-source retrial queueing systems with two-way communications to the orbit and blocking, in: V. M. Vishnevskiy, K. E. Samouylov, D. V. Kozyrev (Eds.), *Distributed Computer and Communication Networks: Control, Computation, Communications*, Springer International Publishing, Cham, 2020, pp. 171–182.
5. V. I. Dragieva, Number of retrials in a finite source retrial queue with unreliable server., *Asia-Pac. J. Oper. Res.* 31 (2) (2014) 23. doi:10.1142/S0217595914400053.
6. A. Tóth, J. Sztrik, A. Pintér, Z. Bács, Reliability analysis of finite-source retrial queueing system with collisions and impatient customers in the orbit using simulation, in: 2021 International Conference on Information and Digital Technologies (IDT), 2021, pp. 230–234. doi:10.1109/IDT52577.2021.9497567.
7. U. C. Gupta, N. Kumar, F. P. Barbhuiya, *A Queueing System with Batch Renewal Input and Negative Arrivals*, Springer Singapore, Singapore, 2020, pp. 143–157. doi:10.1007/978-981-15-5951-8_10.
8. S. R. Chakravarthy, S. Subramanian, A stochastic model for automated teller machines subject to catastrophic failures and repairs, *Industrial & Manufacturing Engineering Publications* 1 (1) (2018) 75–94.
9. P. A. Fishwick, Simpack: Getting started with simulation programming in c and c++, in: In 1992 Winter Simulation Conference, 1992, pp. 154–162.

UDC: 004.85

On Model Inversion Attacks

Junzhe Song¹ and Dmitry Namiot ¹

¹Lomonosov Moscow State University, Moscow, Russia

Abstract

Attacks on machine learning systems are defined as special actions on the elements of the machine learning pipeline, which are designed to either prevent the normal operation of machine learning systems or ensure their special functioning, which is necessary for the attacker. Model inversion attacks aim to expose the private data used to train the model. Attacks that expose private information about machine learning systems are a big threat to machine learning as a service projects. In this article, we provide an overview of off-the-shelf software tools for performing model inversion attacks.

Keywords: adversarial examples, model inversion attacks

1. Introduction

Adversarial Machine Learning becoming now a threat to machine learning security, and is also an area that the software industry needs to focus on. Leaders of the software industry, such as Google [1], IBM [2], and Microsoft [3], declare that except securing their traditional software systems, they will also actively secure ML systems. The leading industry market research company, Gartner, published its first report in Feb 2019 on adversarial machine learning [4], which suggested that "*Application leaders must anticipate and prepare to mitigate potential risks of data corruption, model theft, and adversarial samples.*"

The Model Inversion attack (or MI attack) is a branch of Adversarial Machine Learning. In model inversion attacks, the attacker attempts to recover a private data set used to train a neural network. A successful model inversion attack creates (generates) samples that describe data in a private dataset. In other words, this is an attack that tries to steal the data used during network training. Since the MI attack first entered the sight of the public in 2015, the theories of MI attack have been developed for 7 years. In these years, there are several methods of MI attack have been raised, such as: using confidence informational [5], attacking without

the knowledge of non-sensitive attributes [6], exploiting explanations [7], etc. Many of these methods are only a laboratory version at the time of presentation, Thus, today, several years after these methods were proposed, we are very interested in the implementation of these methods.

The result of the survey of 28 different organizations [8] shows that, despite there being an urge to secure ML systems, most industry practitioners are unready to accept the content of adversarial machine learning. 25 out of the 28 organizations show that they don't have suitable tools to secure their ML systems and are searching for guidance.

Therefore, we collect existing implementations of MI attacks as well as the known protection methods in order to benefit software industry practitioners as well as to determine further directions for the development of software tools for detecting such attacks and combating them.

2. How to implement a model inversion attack?

2.1. Attacks by *Fredrikson et al.*. This is an unofficial implementation. In this project [9], the author implements the MI attack introduced in the paper *Model inversion attacks that exploit confidence information and basic countermeasures* [5]. The author implements this attack with important dependencies **Tensorflow**, **Pylearn2**, **Matplotlib**, and **Scipy**. The target database for the implemented attack is **AT&T Database of Faces** [10].

From the result provided by the author, the effect of the attack is not so well. For example, for an image sample called A , the best image after an inversion attack is a blurred image A_{best} , from which we can hardly tell who he is, perhaps after a higher set of iterations, we can get a clearer attack result than current A_{best} .

Official code can be found at <https://www.cs.cmu.edu/~mfredrik/>.

2.2. AIJack. The **AIJack** is a **Python** library created by Github user *Koukyosyumei* [11] aimed to reveal the vulnerabilities of machine learning models. This package implements algorithms for AI security such as Model Inversion, Poisoning Attack, Evasion Attack, Differential Privacy, and Homomorphic Encryption.

The package contains Collaborative Learning (FedAVG, SplitNN), Attacks (MI-Face[5], Gradient Inversion[12][13][14][15][16], GAN Attack at client-side[17], Label Leakage Attack[18], Evasion Attack[19], Poisoning Attack[20]), and Defense (Moment Accountant(Differential Privacy)[21], DPSGD (Differential Privacy)[21], MID (Defense against model inversion attack)[22], Soteria (Defense against model inversion attack in federated learning)[23]).

From the content above, we can find that this library contains many famous methodologies of attack and defenses for ML models, which means, with the help of this library, freshman of AI security and software industry practitioners can easily create an instance of attack to understand how the adversarial user may attack the system, or create an instance of defense to understand how the current defense strategy defend our system.

The usage of this library is also simple, with several code lines, you can create an attack on a machine learning model. For example, a *MI-FACE* attacker(model inversion attack):

```
1 # Fredrikson attack
2
3 from aijack.attack import MI_FACE
4
5 mi = MI_FACE(target_torch_net, input_shape)
6 reconstructed_data, _ = mi.attack(target_label, lam, num_itr)
```

or a *Gradient Inversion* attacker (server-side model inversion attack against federated learning):

```
1 from aijack.attack import GradientInversion_Attack
2
3 # DLG Attack (Zhu, Ligeng, Zhijian Liu, and Song Han)
4 attacker = GradientInversion_Attack(net, input_shape, distancename="12")
5
6 # GS Attack (Geiping, Jonas, et al.)
7
8 attacker = GradientInversion_Attack(net, input_shape,
9                                     distancename="cossim", tv_reg_coef=0.01)
10
11 # iDLG (Zhao, Bo, Konda Reddy Mopuri, and Hakan Bilen)
12 attacker = GradientInversion_Attack(net, input_shape,
13                                     distancename="12", optimize_label=False)
14
15 # CPL (Wei, Wenqi, et al.)
16 attacker = GradientInversion_Attack(net, input_shape,
17                                     distancename="12", optimize_label=False, lm_reg_coef=0.01)
```

2.3. Privacy Raven. PrivacyRaven is a privacy testing library for deep learning systems [24]. Users can use it to determine the susceptibility of a model to different privacy attacks; evaluate privacy-preserving machine learning techniques; develop novel privacy metrics and attacks; and repurpose attacks for data provenance and other use cases.

PrivacyRaven supports many known attack methods, such as label-only black-box model extraction, membership inference, and model inversion attacks (in a future version). The developer also plans to include differential privacy verification, automated hyperparameter optimization, more classes of attacks, and other features in this library.

Deep learning systems, significantly neural networks, have proliferated in an exceedingly big selection of applications, together with privacy-sensitive use cases like identity verification and medical diagnoses. However, these models are liable to privacy attacks that focus on each intellectual property of the model and the confidentiality of the training data. Recent literature has seen competition between privacy attacks and defenses on varied systems. And until now, engineers and researchers haven't had the privacy analysis tools that match this trend.

Many alternative deep learning security techniques are taxing to use, which discourages their adoption. PrivacyRaven is supposed for a broad audience, so they designed it to be:

- **Usable**-Multiple levels of abstraction permit users to either alter a lot of the interior mechanics or directly manage them, betting on their use case and familiarity with the domain;
- **Flexible**-A standard style makes the attack configurations customizable and practical. It additionally permits new privacy metrics and attacks to be incorporated straightforwardly;
- **Efficient**-PrivacyRaven reduces the boilerplate, affording fast prototyping and quick experimentation.

Each attack may be launched in fewer than 15 lines of code.

2.4. Adversarial Robustness Toolbox. Adversarial Robustness Toolbox(ART) is a Python library for Machine Learning Security [25]. ART provides tools that allow developers and researchers to defend and judge Machine Learning models and applications against the adversarial threats of Evasion, Poisoning, Extraction, and Inference. ART supports all fashionable machine learning frameworks (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.), all types of data (images, tables, audio, video, etc.), and machine learning tasks (classification, object detection, speech recognition, generation, certification, etc.).

The ART contains: Attacks (Evasion Attacks, Poisoning Attacks, Extraction Attacks, Inference Attack), Defenses (Preprocessor, Postprocessor, Trainer, Trans-

former, Detector), Estimators (Classification, Object Detection, Object Tracking, Speech Recognition, Certification, Encoding, Generation), and Metrics (Robustness Metrics, Certification, Verification)

3. On comparison of implementations

Table 1. Comparison of implementations

Implement's name	AIJack	Privacy Raven	ART
Attacks Types			
Evasion Attacks	✓		✓
Inversion Attacks	✓	✓	✓
Poisoning Attacks	✓		✓
Extraction Attacks		✓	✓
Membership Inference Attacks	✓	✓	✓
Label Leakage	✓		
Total of supported	5	3	5
Defences Types			
DPSGD(Differential Privacy)	✓		
MID(Defense against model inversion attack)	✓		
Soteria	✓		
CKKS [27]	✓		
Defensive Distillation [28]			✓
Neural Cleanse [29]			✓
Total of supported	4	0	2

4. Discussion and future work

From table 1 above, we can consider the **ART** as a state-of-the-art tool. And with the support of *IBM* and *DARPA*, the **ART** can grow very quickly. Among many attacks, **ART** can be used to carry out attacks aimed at extracting private data ref30. So far, **ART** includes only two detectors for *Evasion Attacks* and *Poisoning Attacks*, which means the detectors for the rest of the attacks are still under construction. Therefore, our next step is to try to build a detector for *Model Inversion Attacks*, and for this detector, we can start with the *Fredrikson et al.* [5].

This research has been supported by the Interdisciplinary Scientific and Educational School of Moscow University "Brain, Cognitive Systems, Artificial Intelligence". The work was carried out as part of the development of the program of the faculty of the CMC of the Lomonosov Moscow State University "Artificial intelligence in cybersecurity" [31]. For other publications within the framework of this project, see, for example, papers [32] [33].

REFERENCES

1. "Responsible AI Practices." [Online]. Available: <https://ai.google/responsibilities/responsible-ai-practices/?category=security>
2. "Adversarial Machine Learning," Jul 2016. [Online]. Available: <https://ibm.co/36fhajg>
3. "Securing the Future of AI and ML at Microsoft." [Online]. Available: <https://docs.microsoft.com/en-us/security/securing-artificial-intelligence-machine-learning>
4. S. A. Gartner Inc, "Anticipate Data Manipulation Security Risks to AI Pipelines." [Online]. Available: <https://www.gartner.com/doc/3899783>
5. Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1322-1333).
6. Hidano, S., Murakami, T., Katsumata, S., Kiyomoto, S., & Hanaoka, G. (2017, August). Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes. In 2017 15th Annual Conference on Privacy, Security and Trust (PST) (pp. 115-11509). IEEE.
7. Zhao, X., Zhang, W., Xiao, X., & Lim, B. Y. (2021). Exploiting Explanations for Model Inversion Attacks. arXiv preprint arXiv:2104.12669.
8. R. S. Siva Kumar et al., "Adversarial Machine Learning-Industry Perspectives," 2020 IEEE Security and Privacy Workshops (SPW), 2020, pp. 69-75, doi: 10.1109/SPW50608.2020.00028.
9. "Implementation of MI attack with method *Fredrikson et al.*". Available: <https://github.com/yashkant/Model-Inversion-Attack>
10. AT&T Database of Faces. Available: <https://www.kaggle.com/datasets/kasikrit/att-database-of-faces?select=s1>
11. AIJack: reveal the vulnerabilities of machine learning models. Available: <https://github.com/Koukyosyumei/AIJack>
12. Zhu, Ligeng, Zhijian Liu, and Song Han. "Deep leakage from gradients." Advances in Neural Information Processing Systems 32 (2019).
13. Geiping, Jonas, et al. "Inverting gradients-how easy is it to break privacy in federated learning?." Advances in Neural Information Processing Systems 33 (2020): 16937-16947.
14. Zhao, Bo, Konda Reddy Mopuri, and Hakan Bilen. "idlgl: Improved deep leakage from gradients." arXiv preprint arXiv:2001.02610 (2020).
15. Wei, Wenqi, et al. "A framework for evaluating gradient leakage attacks in federated learning." arXiv preprint arXiv:2004.10397 (2020).

16. Yin, Hongxu, et al. "See through gradients: Image batch recovery via gradient-version." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
17. Hitaj, Briland, Giuseppe Ateniese, and Fernando Perez-Cruz. "Deep models under the GAN: information leakage from collaborative deep learning." Proceedings of the #2017 ACM SIGSAC Conference on Computer and Communications Security. 2017.
18. Li, Oscar, et al. "Label leakage and protection in two-party split learning." arXiv preprint arXiv:2102.08504 (2021).
19. Biggio, Battista, et al. "Evasion attacks against machine learning at test time." Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2013.
20. Biggio, Battista, Blaine Nelson, and Pavel Laskov. "Poisoning attacks against support vector machines." arXiv preprint arXiv:1206.6389 (2012).
21. Abadi, Martin, et al. "Deep learning with differential privacy." Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016.
22. Wang, Tianhao, Yuheng Zhang, and Ruoxi Jia. "Improving robustness to model inversion attacks via mutual information regularization." arXiv preprint arXiv:2009.05241 (2020).
23. Sun, Jingwei, et al. "Soteria: Provable defense against privacy leakage in federated learning from representation perspective." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
24. Privacy Raven: Privacy Testing for Deep Learning. [Online]. Available: <https://github.com/trailofbits/PrivacyRaven>
25. Adversarial Robustness Toolbox (ART) - Python Library for Machine Learning Security - Evasion, Poisoning, Extraction, Inference - Red and Blue Teams. Available: <https://github.com/Trusted-AI/adversarial-robustness-toolbox>
26. Advertorch: A Toolbox for Adversarial Robustness Research. [Online]. Available: <https://github.com/borealisai/advertorch>
27. Cheon, J. H., Kim, A., Kim, M., & Song, Y. (2017, December). Homomorphic encryption for arithmetic of approximate numbers. In International Conference on the Theory and Application of Cryptology and Information Security (pp. 409-437). Springer, Cham.
28. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, May). Distillation as a defense to adversarial perturbations against deep neural networks. In 2016 IEEE symposium on security and privacy (SP) (pp. 582-597). IEEE.
29. Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., & Zhao, B. Y. (2019, May). Neural cleanse: Identifying and mitigating backdoor attacks in

- neural networks. In 2019 IEEE Symposium on Security and Privacy (SP) (pp. 707-723). IEEE.
- 30. Attacks against Machine Learning Privacy (Part 1): Model Inversion Attacks with the IBM-ART Framework [Online]. Available: <https://franziska-boenisch.de/posts/2020/12/model-inversion/>
 - 31. Artificial Intelligence in Cybersecurity. [Online]. Available: <http://master.cmc.msu.ru/?q=ru/node/3496> (in Russian) Retrieved: Jul, 2022.
 - 32. Dmitry, Namiot, Ilyushin Eugene, and Chizhov Ivan. "On a formal verification of machine learning systems." International Journal of Open Information Technologies 10.5 (2022): 30-34.
 - 33. Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilipenko. "On Trusted AI Platforms." International Journal of Open Information Technologies 10.7 (2022): 119-127. (in Russian).

УДК: 519.872

Метод расчета характеристик системы обслуживания с гибридной политикой пополнения запасов от двух источников

А.З. Меликов¹, Р.Р. Мирзоев², С.С. Наир³

¹Институт Систем Управления, Национальная Академия Наук Азербайджан,
Баку, Азербайджан

²Национальная Авиационная Академия, Баку, Азербайджан

³Государственный инженерный колледж, Триссур, Индия

agassi.melikov@gmail.com

Аннотация

Предложена гибридная политика пополнения запасов (ППЗ) в системах обслуживания-запасания с двумя источниками. Если уровень запасов опускается до точки заказа s , то используется (s, Q) политика, при этом пополнения запасов осуществляется из медленного и недорогого источника. Если уровень запасов опускается ниже определенной пороговой величины r , $r < s$, то используется (r, S) политика, где пополнения запасов осуществляется из быстрого и дорогого источника. В систему кроме расходящих заявок поступают еще и разрушающие заявки, которые не требуют запасов, а разрушают их. Найдено условие эргодичности исследуемой системы, вычисляется ее стационарное распределение и предложены формулы для нахождения ее характеристик.

Ключевые слова: система обслуживания-запасания, гибридная политика пополнения запасов, два источника поставок, матрично-геометрический метод.

1. Введение

В подавляющем большинстве работ рассматриваются СОЗ с одним источником пополнения запасов, при этом считается во всем периоде работы используется единая политика пополнения запасами (ППЗ). Однако в целях увеличения надежности своевременного обслуживания расходящих заявок (c -заявок) необходимо организовать поставки запасов из нескольких источников. Одна из моделей СОЗ данного типа рассмотрена в [1], в которой решена проблема распределения заказов между быстрым и дорогим поставщиком и медленным, но недорогим поставщиком.

Научный и практический интерес представляют изучение моделей СОЗ, в которых в зависимости от текущей ситуации используются различные ППЗ. Такие ППЗ будем называть гибридными. Несмотря на их важность, модели СОЗ с несколькими источниками поставок и гибридной ППЗ почти не исследованы. Подобная модель СОЗ с гибридной ППЗ и с одним источником поставок изучена в [2]. В ней рассмотрена СОЗ с портящимися запасами, ограниченным буфером для первичных заявок и неограниченным буфером для заявок, образуемых в результате обратной связи. Считается, что в зависимости от уровня запасов системы используется либо ППЗ с фиксированным, либо с переменными объемами поставок. В работе [3] изучена модель СОЗ с портящимися запасами, гибридной ППЗ и двумя источниками. В ней предложена следующая гибридная ППЗ: когда уровень запасов опускается до фиксированной величины m , $m > (S/2)$, срабатывает обычный заказ объема $Q_1 = S - m$ к медленному источнику; если уровень запасов опускается до заранее определенной величины s , $s < Q_1$, срабатывает экстренный заказ объема $Q_2 = S - s > s + 1$ к быстрому источнику. С помощью матрично-геометрического метода [4] найдено стационарное распределение соответствующей цепи Маркова (ЦМ) и решена задача минимизации суммарных штрафов системы.

В недавней работе [5] рассмотрены модели СОЗ с двумя источниками, в которых возможны мгновенные порчи запасов из-за внезапных событий. Настоящая работа является продолжением исследований, начатых в [5]. Здесь предложена гибридная ППЗ в системах с возможностью мгновенной порчи запасов, двумя источниками поставок и аннулированием заказа от медленного источника.

2. Описание модели и предложенной политики

Интенсивность пуассоновского потока c -заявок равна λ , при этом каждая c -заявка требует запаса единичного размера. Поток разрушающих заявок (d -заявок) также является пуассоновским с параметром κ , при этом в момент поступления таких заявок уровень запасов мгновенно уменьшается на единицу; d -заявка может даже уничтожить запас, который находится на этапе отпуска к c -заявке. Если в момент поступления c -заявки сервер свободен и уровень запасов положительный, то она немедленно принимается для обслуживания; если уровень запасов положительный и сервер занят, то c -заявка становится в очередь бесконечной длины. Если в момент поступления c -заявки уровень запасов равен нулю, то она либо с вероятностью φ_1 становится в очередь, либо с дополнительной вероятностью $\varphi_2 = 1 - \varphi_1$ покидает систему. Если до начала обслуживания c -заявки уровень запасов падает до нуля, то c -заявка во главе очереди покидает систему с неудовлетворенным спросом после некоторое случайное время, которое имеет показательную ф.р. со средним τ^{-1} .

По завершении обслуживания *c*-заявка она либо с вероятностью σ_1 не покупает товар, либо с дополнительной вероятностью $\sigma_2 = 1 - \sigma_1$ покупает товар. В обоих случаях время обслуживания *c*-заявок имеют показательные ф.р., при этом если *c*-заявка отказывается купить товар, то среднее время ее обслуживания равно μ_1^{-1} ; иначе это время равно μ_2^{-1} .

Пополнения запасов производятся из двух источников: медленного Источника-1 и быстрого Источника-2. Время выполнения заказов от каждого источника имеет показательную ф.р., при этом время выполнения заказа от Источнику-*i* имеет экспоненциальную ф.р. со средним ν_i^{-1} , $i = 1, 2$, при этом $\nu_2 > \nu_1$.

Если уровень запасов опускается до значение s , $0 < s < (S/2)$, то делается заказ объема $Q = S - s$ к Источнику-1, а когда уровень запасов опускается до пороговой величины r , $0 \leq r < s$, то мгновенно аннулируется заказ от Источника-1 и отправляется заказ объема $S - r$ к Источнику-2, при этом количество пополнения должно быть таким, чтобы вернуть уровень запасов к S в эпоху пополнения.

Задача состоит в нахождении совместного распределения числа *c*-заявок в системе и уровня запасов системы, а также вычислении основных характеристик системы.

3. Расчет вероятностей состояний системы

Работа системы описывается двумерной ЦМ с состояниями вида (n, m) , где n указывает число *c*-заявок в системе, $n \geq 0$, m обозначает уровень запасов на складе системы, $m = 0, 1, \dots, S$. Пространство состояний (ПС) этой ЦМ определяется так:

$$E = \bigcup_{n=0}^{\infty} L(n),$$

где $L(n) = \{(n, 0), (n, 1), \dots, (n, S)\}$ – n -й уровень.

Интенсивность перехода из состояния $(n_1, m_1) \in E$ в другое состояние $(n_2, m_2) \in E$ обозначим через $q((n_1, m_1), (n_2, m_2))$. Положительные элементы генератора изучаемой цепи определяются так:

$$(n_1, m_1), (n_2, m_2)) = \begin{cases} \lambda\varphi_1, & \text{если } n_2 = n_1 + 1, m_2 = m_1 = 0, \\ \lambda, & \text{если } n_2 = n_1 + 1, m_2 = m_1 > 0, \\ \mu_1\sigma_1, & \text{если } n_2 = n_1 - 1, m_2 = m_1 > 0, \\ \mu_2\sigma_2, & \text{если } n_2 = n_1 - 1, m_2 = m_1 - 1, \\ \kappa, & \text{если } n_2 = n_1, m_1 > 0, m_2 = m_1 - 1, \\ \tau, & \text{если } n_1 > 0, n_2 = n_1 - 1, m_2 = m_1 = 0, \\ v_1, & \text{если } n_2 = n_1, r < m_1 \leq s, m_2 = m_1 + S - s, \\ v_2, & \text{если } n_2 = n_1, 0 \leq m_1 \leq r, m_2 = S. \end{cases}$$

Из соотношений (1) видно, что полученная двумерная ЦМ описывается квазипроцессом размножения и гибели, в который интенсивности переходов не зависят от уровня (Level Independent Quasi-Birth-Death (LIQBD) Process). Перенумеровав состояния данной двумерной ЦМ лексикографическим способом (т.е. состояния нумеруются согласно порядку $(0, 0), (0, 1), \dots, (0, S), (1, 0), (1, 1), \dots, (1, S), \dots$), находим, что генератор полученного LIQBD имеет следующий вид:

$$G = \begin{pmatrix} B & A_0 & O & \dots & O & \dots \\ A_2 & A_1 & A_0 & O & O & \dots \\ O & A_2 & A_1 & A_0 & O & \dots \\ O & O & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \end{pmatrix},$$

где O означает нулевую квадратную матрицу размерности $S + 1$, а блочные матрицы $B = \|b_{ij}\|$, $A_k = \|a_{ij}^{(k)}\|$, $k = \overline{0, 2}$, $i, j = \overline{0, S}$, являются квадратными с той же размерностью, где их ненулевые элементы определяются как:

$$b_{ij} = \begin{cases} \nu_2, & \text{если } 0 \leq i \leq r, j = S, \\ \nu_1, & \text{если } r < i \leq s, j = i + S - s, \\ \kappa, & \text{если } 0 < i \leq S, j = i - 1, \\ -(\nu_2 + \lambda\varphi_1), & \text{если } i = j = 0, \\ -(\nu_2 + \kappa + \lambda), & \text{если } 0 < i \leq r, i = j, \\ -(\nu_1 + \kappa + \lambda), & \text{если } r < i \leq s, i = j, \\ -(\kappa + \lambda), & \text{если } s < i \leq S, i = j; \end{cases} \quad (1)$$

$$a_{ij}^{(0)} = \begin{cases} \lambda\varphi_1, & \text{если } i = j = 0, \\ \lambda, & \text{если } i > 0, i = j; \end{cases} \quad (2)$$

$$a_{ij}^{(1)} = \begin{cases} \nu_2, & \text{если } 0 \leq i \leq r, j = S, \\ \nu_1, & \text{если } r < i \leq s, j = i + S - s, \\ \kappa, & \text{если } 0 < i \leq S, j = i - 1, \\ -(\tau + \nu_2 + \lambda\varphi_1), & \text{если } i = j = 0, \\ -(\nu_2 + \kappa + \lambda + \mu_1\sigma_1 + \mu_2\sigma_2), & \text{если } 0 < i \leq r, i = j, \\ -(\nu_1 + \kappa + \lambda + \mu_1\sigma_1 + \mu_2\sigma_2), & \text{если } r < i \leq s, i = j, \\ -(\kappa + \lambda + \mu_1\sigma_1 + \mu_2\sigma_2), & \text{если } s < i \leq S, i = j; \end{cases} \quad (3)$$

$$a_{ij}^{(2)} = \begin{cases} \tau, & \text{если } i = j = 0, \\ \mu_1\sigma_1, & \text{если } i > 0, i = j, \\ \mu_2\sigma_2, & \text{если } i > 0, j = i - 1. \end{cases} \quad (4)$$

Теорема. Система является эргодичной тогда и только тогда, когда выполняется следующее соотношение:

$$\lambda(1 - \varphi_2\pi(0)) < \tau\pi(0) + (\mu_1\sigma_1 + \mu_2\sigma_2)(1 - \pi(0)), \quad (5)$$

$$\text{где } \pi(0) = \left(\sum_{m=0}^{r+1} \alpha_m + \sum_{m=r+2}^{s+1} \beta_m + (S - 2s + r - 1) \beta_{s+1} + \sum_{m=S-s+r+1}^S \gamma_m \right)^{-1};$$

$$\alpha_m = \begin{cases} 1, & \text{если } m = 0 \\ \theta_2(1 + \theta_2)^{m-1}, & \text{если } 1 \leq m \leq r+1; \end{cases}$$

$$\beta_m = \theta_2 \left(\frac{1 + \theta_2}{1 + \theta_1} \right)^r (1 + \theta_1)^{m-1}, \quad r+1 < m \leq s+1;$$

$$\gamma_m = \theta_2 \sum_{k=0}^r \alpha_k + \theta_1 \sum_{k=m-(S-s)}^s \beta_k, \quad S-s+r+1 < m \leq S; \quad \theta_i = \nu_i / (\mu_2\sigma_2 + \kappa), \quad i = 1, 2.$$

Согласно алгоритму для LIQBD (см. [4, с. 81-83]), заключаем, что при выполнении условия эргодичности (5) стационарное распределение $p = (p_0, p_1, p_2, \dots)$, $p_n = (p(n, 0), p(n, 1), \dots, p(n, S))$, соответствующее генератору G , определяется как

$$p_n = p_0 R^n, \quad n \geq 1,$$

где R является неотрицательным и минимальным решением следующего квадратичного матричного уравнения:

$$R^2 A_2 + R A_1 + A_0 = 0.$$

Вероятности p_0 граничных состояний вычисляются из СУР:

$$p_0(B + R A_2) = 0,$$

$$p_0(I - R)^{-1} e = 1,$$

где I обозначает единичную матрицу размерности $S + 1$.

4. Расчет характеристик системы

Характеристики системы находятся с помощью вероятностей состояний системы следующим образом.

средний уровень запасов на складе (S_{av})

$$S_{av} = \sum_{m=1}^S m \sum_{n=0}^{\infty} p(n, m);$$

средний объем поставок от Источника- i , $i = 1, 2, (V_{av}(i))$

$$V_{av}(1) = (S - s) \sum_{m=r+1}^s \sum_{n=0}^{\infty} p(n, m) ; V_{av}(2) = \sum_{m=0}^r (S - m) \sum_{n=0}^{\infty} p(n, m) ;$$

среднее число c -заявок в системе (L_{av})

$$L_{av} = \sum_{n=1}^{\infty} n \sum_{m=0}^S p(n, m) ;$$

средняя интенсивность уничтожения запасов системы (DRS):

$$DRS = \kappa \left(1 - \sum_{n=0}^{\infty} p(n, 0) \right) ;$$

средняя интенсивность обычных заказов (RR_1):

$$RR_1 = \kappa p(0, s + 1) + (\mu_2 \sigma_2 + \kappa) \sum_{n=1}^{\infty} p(n, s + 1) ;$$

средняя интенсивность экстренных заказов (RR_2):

$$RR_2 = \kappa p(0, r + 1) + (\mu_2 \sigma_2 + \kappa) \sum_{n=1}^{\infty} p(n, r + 1) ;$$

вероятность потери заявок c -заявок (PL);

$$PL = \varphi_2 \sum_{n=0}^{\infty} p(n, 0) + \frac{\tau}{\tau + \lambda \varphi_1 + v_2} \sum_{n=1}^{\infty} p(n, 0) ;$$

5. Заключение

Предложена гибридная ППЗ в системах обслуживания-запасания с бесконечной очередью и двумя источниками поставок, где быстрый источник является более дорогим, чем медленный источник. Она основана на комбинированном использовании двух известных ППЗ – с фиксированным объемом поставок и с переменным объемом поставок. Другая отличительная особенность изучаемой модели состоит в том, что порчи запасов происходят мгновенно в моменты поступления разрушающих заявок. Построена математическая модель изучаемой системы в виде двумерной ЦМ, которая имеет трехдиагональный генератор. Найдено условие эргодичности полученной ЦМ и использован матрично-геометрический метод для нахождение ее стационарное распределение. Предложены формулы для вычисления характеристик изучаемой системы.

Литература

1. Melikov A., Krishnamoorthy A., Shahmaliyev M.O. Numerical analysis and long run total cost optimization of perishable queuing inventory systems with delayed feedback // Queuing Models and Service Managements. 2019. V. 2. Iss. 1. P. 83–111.
2. Amirthakodi M., Radhamani V., Sivakumar B. A perishable inventory system with service facility and feedback customers // Annals of Operations Research. 2015. V. 233. P. 25-55.
3. Soujanya M.L., Laxmi P.V. Analysis on dual supply inventory model having negative arrivals and finite lifetime inventory // Reliability: Theory and Applications. 2021. V. 16. № 3. P. 295–301.
4. Neuts M.F. Matrix-geometric solutions in stochastic models: An algorithmic approach. Baltimore: John Hopkins University Press, 1981.
5. Melikov A.Z., Mirzayev R.R., Nair S.S. Numerical investigation of queuing-inventory system with double sources and destructive customers // Journal of Computer and System Science International. 2022. V. 61. Iss. 4.

UDC: 004.85

On monitoring of machine learning models

Dmitry Namiot¹ and Eugene Ilyushin¹

¹Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russian Federation

dnamiot@gmail.com, john.ilyushin@gmail.com

Abstract

In this article, we want to focus on monitoring machine learning models. The practical use of machine learning systems includes several fairly standard steps that do not depend on the subject area. These stages of the pipeline are described in sufficient detail, all descriptions of machine learning systems begin with them and include steps such as preparing the model, selecting training data, training the model, and testing it (putting it into operation). At the same time, such a stage as monitoring during the operation phase is almost always excluded from consideration. Meanwhile, this moment plays a key role, for example, to ensure the stability of the machine learning system. This article discusses the practical issues of building monitoring for machine learning systems.

Keywords: monitoring, data shift, concept shift

1. Introduction

The development and operation of machine learning systems are inextricably linked to the so-called machine learning pipeline. Any description of such systems begins with a description of the steps of this pipeline. The names of the steps may vary [1], but, in general, the following stages can be distinguished:

- Model building. This refers to the analysis of the problem, understanding of the availability of data, the choice of architecture and algorithms, and the selection of training and test data.
- Model training. Performance evaluation, model performance, hyperparameters tuning
- Model testing. Confirmation of the characteristics obtained during the training phase.

And after that, there are two "production" stages that are often skipped, but that doesn't make them any less important. Firstly, it is the preparation and launch of the industrial (production) version of the developed system [2]. Its implementation can be done on completely different platforms (and even programming languages), for example. And the last stage is monitoring. This is already a continuous process during the operation of the system, which, in short, should keep track of the fact that the system does (continues to do) what was expected of it in the previous step. It is this process of monitoring that is the subject of this article. Conceptually, the whole model looks very simple and logical [3].

Monitoring results are the basis for revising decisions at different stages of the life cycle. However, we have deliberately marked these references with a question mark, as such a change may be difficult to implement in practice. For example, our machine learning system runs 24/7 as a critical application. What does it mean, in practice, to retrain the model (return to the training phase in Figure 1)? Will the system be interrupted indefinitely? Some papers seriously discuss the need for a quick reaction of system developers to possible problems [4], but for systems that must work continuously, a quick reaction will obviously not help, since interruptions in work are simply unacceptable. It should be noted that it is critical applications that are critical in terms of the importance of monitoring. The continuity of these sorts of systems is a major constraint on any return to the previous stages of the machine learning pipeline – remodeling and retraining.

The rest of the article is structured as follows. In section II, we consider the objects of monitoring. Section III focuses on existing tools and a possible pipeline.

2. What should be monitored for machine learning systems?

The purpose of monitoring is to help resolve possible problems at the stage of operation of the machine learning system. Monitoring should provide data on the basis of which the solution to emerging problems will be carried out. At the same time, it should be noted that the monitoring of machine learning systems is fundamentally different from the monitoring of "traditional" software systems. The reason is obvious - the non-deterministic nature of algorithms and conclusions. Basically, all machine learning systems (models) are trained on some subset of data (training set). It is always some subset of the (often unknown) general population. If we are talking about discriminant systems here (and all systems for critical use are just such), then further work, for example, of a trained classifier is based only on the fact that the real data (general population) do not differ (do not differ much) from the training ones. If the data is different, then all the generalizations achieved during the training phase may turn out to be incorrect. This is the first and main difference from the monitoring of "traditional" software systems. It is necessary to track the possible

change in real data compared to the training dataset. This is a separate and big question about what to do if such a difference is found, but as a first step, it must be found.

This difference is called drift. In this regard, it is customary to talk about data shift (data drift) and concept shift (concept drift) [5]. Data drift corresponds to the situation when the distribution of input data changes compared to the one that was at the training stage of the model. In other words, the real input data looks somehow different. Concept drift corresponds to the situation when the relationships (rules, connections) between the input variables and the output fundamentally change.

It is necessary to dwell on the last concept separately. From the point of view of a formal representation, of course, it is good to have a single concept of drift. But from an engineering point of view, this is a kind of disaster. What if the equipment used is set to measure specific parameters that were used in the model to obtain a response (classification), and now it turns out that the measurement data does not determine the response, etc.? In practice, this is a statement that a new model is needed. And if data collection is tied to any hardware solutions, then, accordingly, new equipment. Therefore, in the existing works, the conceptual shift is considered as a change in the joint distribution. That is, both inputs and outputs remain unchanged. As a typical example of a conceptual shift, consider the example of a forecasting system that has experienced a significant change in trend. For example, ticket sales forecasts for offline events and restrictions related to COVID. The pattern of consumer behavior has changed dramatically.

The next item in monitoring is, of course, performance. But here we need to keep in mind that we may not have data to evaluate, for example, the accuracy of the classifier. There is input data, there is the result of the model, but there is no true result. And, of course, performance evaluation does not replace data shift analysis. If only because the first reason to explain the change in performance is data drift.

And the third point should be the quality of the data. In particular: the presence of gaps in data, respect for the ranges of values for variables, and the absence of new (unknown) categorical data.

In our further consideration:

X is a feature space,

Y is a label space

$P(X)$: is a distribution of features

$P(Y)$: is a distribution of labels

$P(X | Y)$: is a distribution of features given specific labels

$P(Y | X)$: is a distribution of labels given specific features

$P(X, Y)$: joint distribution

This is what ML models are trying to discover. The various drifts are defined as follows:

Covariate shift

$P(Y | X)$ is the same but $P(X)$ changes

Label shift

$P(Y)$ changes but $P(X | Y)$ is the same

Concept shift

$P(Y | X)$ changes but $P(X)$ is the same

Some papers talk about joint distribution.

Concept drift - joint distribution changes over time

$P_t(X, Y) \neq P_{t1}(X, Y)$

Sometimes these groups may be called differently, although the meaning of the division remains:

Covariate Shift: a shift in the independent variables.

Prior Probability Shift: a shift in the target variable.

Concept Drift: a change (shift) of the relation (connection) between the input (independent) variables and the output.

There is also a group of tasks (processes) associated with the period of operation of machine learning systems. Some of them are combined under the general name technical debt [6]. In addition to the above, academic articles, manuals, and reviews list the following.

Additional Feature Requirement. Customers may request additional features after deployment or even the in-service development team may decide to add new options. For example, to tune software performance, fix bugs, etc.

Changes in modeling. The developer's team may realize that changes in model architecture are beneficial.

Changes in the data processing. The developer's team may realize that changing some of the data processing scripts is beneficial.

Some unforeseen issues. Most often - the emergence of a new category of data.

This is a separate topic related to DevOps tools, frameworks, and software engineering for machine learning systems.

3. About monitoring models

The first obvious idea for determining data drift is to compare the real (working) and training datasets. It means mapping the data from features in the training data to their counterparts in the real (production) data and run the different statistical tests depending on the types of features.

A classic example is Evidently AI [7]. The following types of statistical tests are supported in the system, depending on the types of data being compared:

- Numeric features - a two-sample Kolmogorov-Smirnov test. It is checked whether the test and training data belong to the same distribution or if there is a statistically significant difference.
- Binary categorical features: a simple Z-test for a difference in proportions. How often do the training and test data have one of the two values for the binary feature? This test should verify if there is a statistically significant difference.
- Multivariate categorical features: a chi-squared test. This test aims to see if the distribution of the feature in the test (real) data is likely based on the distribution in the training data.

This approach is based on approaches that have long been used in mathematical statistics. But from a practical point of view, questions, of course, remain. Firstly, collecting a real dataset takes time, this may not be acceptable for critical applications. An open question is about the size of the real dataset, as well as what to do with non-stationary processes. In other words, distributions can differ only over a certain time interval (or intervals). And it is natural that in practical problems we deal with streams, not data sets [12].

Another example is the Alibi Detect package [8, 9]. It adds significantly more statistical methods. For example, Fisher's Exact Test, Maximum Mean Discrepancy (MMD), et al. For example, the context-aware maximum mean discrepancy drift detector [10] is a kernel based method for detecting drift that can take the relevant context into account. Simply, this method allows you to take into account the "permissible" differences between the real and training sets, binding them to some context.

An example of relatively simple logging is mltrace [11].

But most of the presented methods target a covariate shift. Whereas the most "dangerous" is precisely the drift of the concept, since it partially or even completely

denies the existing machine learning model. Technically, the following methods can be used to deal with concept shift.

First, it is online learning where the model is updated on the fly as the model processes one portion of data at a time. In practice, the majority of the critical applications run on streaming data. So, at least theoretically, online learning should be the most natural way to deal with concept drift. But there is one big problem here. These are adversarial examples that are much more difficult to determine when stepping through.

Secondly, it can be periodic retraining of the model. The reason for retraining may be a drop in performance (e.g. accuracy) below some predetermined level. The problem here is the need to evaluate (mark up) real data and the actual retraining. The solution may be to support the operation of a parallel system (double), on which training is performed. Retraining can be carried out on some representative samples.

Another option is to use an ensemble of models. And evaluate the output as a weighted average of the outputs of the individual models.

And perhaps one of the simplest solutions, although not guaranteed to succeed, is ablation. Feature dropping is a well-known way to deal with the concept drift [13]. Actually, there are two ablation-based approaches: feature ablation and model ablation. The goal of a feature ablation is to understand the influence of the different features of the dataset on the performance of the system. A typical example is the package LOFO [14]. It evaluates the performance of the model with all the input features included, then iteratively removes one feature at a time, retrains the model, and evaluates its performance on a validation set. A model ablation involves removing or changing components of the model followed by training and system evaluation using these variations of the model.

4. Conclusion

It should be noted that there is no exhaustive (generally used, covering all areas) monitoring system. The main open issues for today are the following.

Analysis of streaming data. For critical applications, it may not be possible to collect a dataset for comparison with the training set. This requires time, during which the system, if anything, will already work incorrectly

The incremental definition of the shift remains an open question. In practical systems (especially for critical applications), we are always talking about processing streams.

Incremental learning schemes will be sensitive to adversarial examples.

One of the promising areas, in our opinion, is the creation of digital twins of machine learning systems. It is on the twin that it is necessary to work out a new

model, which is then transferred to the industrial system.

The article is a continuation of a series of publications dedicated to sustainable machine learning models [15, 16, 17]. It was prepared as part of the project of the Department of Information Security of the Faculty of Computer Science of Lomonosov Moscow State University on the creation and development of the master's program "Artificial Intelligence in Cybersecurity" [18].

This research has been supported by the Interdisciplinary Scientific, Educational School of Moscow University "Brain, Cognitive Systems, Artificial Intelligence"

REFERENCES

1. Zöller, Marc-André, and Marco F. Huber. "Benchmark and survey of automated machine learning frameworks." *Journal of artificial intelligence research* 70 (2021): 409-472.
2. Baier, Lucas, Fabian Jöhren, and Stefan Seebacher. "Challenges in the Deployment and Operation of Machine Learning in Practice." ECIS. 2019.
3. Lwakatare, Lucy Ellen, Ivica Crnkovic, and Jan Bosch. "DevOps for AI—Challenges in Development of AI-enabled Applications." 2020 International Conference on Software, Telecommunications and Computer Networks (Soft-COM). IEEE, 2020.
4. Wu, Yujun, Edgar Dobriban, and Susan Davidson. "DeltaGrad: Rapid retraining of machine learning models." International Conference on Machine Learning. PMLR, 2020.
5. Webb, Geoffrey I., et al. "Analyzing concept drift and shift from sample data." *Data Mining and Knowledge Discovery* 32.5 (2018): 1179-1199.
6. Huang, Chong, Arash Nourian, and Kevin Gries. "Hidden Technical Debts for Fair Machine Learning in Financial Services." arXiv preprint arXiv:2103.10510 (2021).
7. Evidently AI <https://evidentlyai.com/> Retrieved: May, 2022
8. Klaise, Janis, et al. "Monitoring and explainability of models in production." arXiv preprint arXiv:2007.06299 (2020).
9. Alibi Detect <https://github.com/SeldonIO/alibi-detect> Retrieved: May, 2022
10. Cobb, Oliver, and Arnaud Van Looveren. "Context-Aware Drift Detection." arXiv preprint arXiv:2203.08644 (2022).
11. Mltrace <https://mltrace.readthedocs.io/en/latest/> Retrieved: May, 2022
12. Mangat, Veenu, Vishal Gupta, and Renu Vig. "Methods to investigate concept drift in big data streams." *Knowledge Computing and its Applications*. Springer, Singapore, 2018. 51-74.

13. Sheikholeslami, Sina. "Ablation programming for machine learning." (2019).
14. LOFO <https://github.com/aerdem4/lofo-importance> Retrieved: May, 2022
15. Li, Huayu, and Dmitry Namiot. "A Survey of Adversarial Attacks and Defenses for image data on Deep Learning." International Journal of Open Information Technologies 10.5 (2022): 9-16.
16. Namiot, D., Ilyushin, E., & Chizhov, I. (2021). The rationale for working on robust machine learning. International Journal of Open Information Technologies, 9(11), 68-74. (in Russian)
17. Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Ongoing academic and industrial projects dedicated to robust machine learning." International Journal of Open Information Technologies 9.10 (2021): 35-46. (in Russian)
18. Artificial Intelligence in Cybersecurity. <http://master.cmc.msu.ru/?q=ru/node/3496> (in Russian) Retrieved: Apr, 2022

UDC: 519.872, 519.217

Queuing system with threshold-based general renovation mechanism

Viana C. C. Hilquias¹, I. S. Zaryadov^{1,2}, T. A. Milovanova¹

¹Department of Applied Probability and Informatics, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

²Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

hilvianamat1@gmail.com, zaryadov-is@rudn.ru, milovanova-ta@rudn.ru

Abstract

The paper considers a single-line queuing system with a single-threshold mechanism for probabilistic dropping of applications accepted into the system (general renovation). On the one hand, unlike the previously considered systems with general renovation, this system introduces a certain threshold value in the queue as a control parameter of the renovation mechanism, which not only determines the moment when the probabilistic dropping of applications accepted into the system is enabled, but also sets a safe area in the queue from which applications accepted into the system cannot be reset. A general renovation is a probabilistic reset of an arbitrary number of applications from the queue outside the safe zone, which occurs at the end of the application service on the device. For this system, the main probabilistic-time characteristics are obtained.

Keywords: queuing system, renovation mechanism, probabilistic-time characteristics

1. Introduction

It should be noted that all AQM algorithms [1, 2, 3, 4, 5, 6, 7] are usually based on a rule for intelligent dropping of packets from the buffer (queue) as its level increases. Unlike RED type AQMs [1, 5] (when the possible packets dropping occurs at the arrival times and the control parameter(s) depending on the queue length)

This paper has been supported by the RUDN University Strategic Academic Leadership Program (Viana C. C. Hilquias, T.A. Milovanova and I.S. Zaryadov, mathematical model formulation and simulation model development). Also the publication has been funded by Russian Foundation for Basic Research (RFBR) according to the research project No. 20-07-00804 (T.A. Milovanova and I.S. Zaryadov, mathematical model development and numerical analysis).

the renovation-base AQM is formed on completely different idea: the decision about a possible dropping is synchronized with the service completions (see [8, 9, 10, 11, 12, 14]). Here we elaborate further on the mechanism of renovation and by analogy with most AQM algorithms, a control parameter was introduced into the renovation mechanism [12, 13] — a certain threshold value in the queue, upon overcoming which the renovation mechanism is activated (and, depending on the model, this threshold value could also limit the area in the drive from which received applications were not dropped).

This work continues the study of systems with renovation and general renovation, first considered in [8, 9, 10]. In [12], it was proposed to use renovation and general renovation as active queue management mechanisms (the development of modern versions of which is still an urgent task [4, 5]), and in [7, 11] the comparison was made of the main characteristics for models with renovation and models with traditional active queue management algorithms (the classic random early detection (RED) [1]).

The key difference of the presented work from the previous ones [12, 13] is that threshold control of the general renovation mechanism is introduced here. It is worth noting that the threshold models with renovation have already been considered by the authors, but for the case when a customer ending its service on the device (at the moment immediately before leaving the device and the system) can drop all customers from the buffer (full renovation or renovation), and not a fixed number of applications with some given probability.

Structurally, the work consists of the following sections: section 2 presents the main results for the queuing system under the first setting, in the section 3 the stationary distribution of the number of applications in the system is presented, the section 4 is devoted to probabilistic characteristics of the system, the section 5 presents time characteristics and the last section 6 summarizes the results.

2. System description

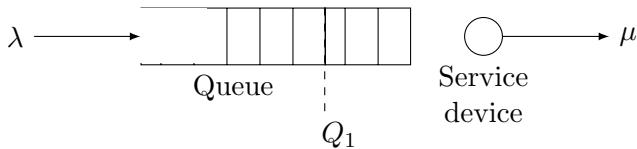


Fig. 1. Diagram of the system with threshold and renovation

Consider a queuing system consisting of one maintenance device and an unlimited capacity storage device. The system receives a recurrent flow of applications ($A(x)$ — the time distribution function between the receipt of applications), the service on the

device is subject to the exponential law with the parameter μ . The threshold value Q_1 in the queue is also determined: if the number of applications in the queue is greater than Q_1 , the application finishing service on the device can reset an arbitrary k ($k \geq 0$) number of applications from the queue with a given probability of dropping $q(k)$ and leave the system.

We will assume that applications are serviced and dropped in the order of receipt.

The study (as in other works on this topic) will be carried out using the embedded Markov chain at the moment of receipt.

Transition probabilities $p_{i,j}, i \geq 0, j \geq 0$ (the probability that at the time of receipt the application will find j applications in the system if the previous received application found i applications in the system) of the embedded Markov chain will be further either denoted, or calculate by using probabilities $A_{i-j+1}^{(k)}, i \geq 0, j \geq 0, k = 1, 2, 3, 4$ — probabilities that between successive arrivals of applications to the system, exactly $i + j + 1$ applications will leave the system (will be serviced and/or will be dropped).

$$p_{i,i+1} = A_0 = \int_0^\infty e^{-\mu x} dA(x) = \alpha(\mu). \quad (1)$$

For the system under consideration, the options are:

- **a)** $0 \leq i \leq Q_1$ (at the time of receipt of the application, the threshold value Q_1 has not been overcome).
- **b)** $i > Q_1$ (taking into account the received application, the threshold value Q_1 is overcome): **b.1)** $j > Q_1 + 1$ (by the time of a new application arrival, the threshold value will be overcome); **b.2)** $j = Q_1 + 1$ (by the time of a new application arrival, the threshold will be reached, but not overcome); **b.3)** $j = Q_1$; **b.4)** $0 \leq j < Q_1$ (by the time of a new application arrival, the threshold is not overcome).

We introduce auxiliary probabilities $\pi_k(l)$ — the probability that k served applications will drop exactly l other applications from the queue ($l \geq 0, k \geq 0$).

$$\pi_1(l) = \begin{cases} q(l), & l > 1, \\ q(0) = p, & l = 0, \end{cases} \quad \pi_k(l) = \sum_{i=0}^l \pi_1(i) \pi_{k-1}(l-i), \quad k > 1, l \geq 0. \quad (2)$$

and $\tilde{\pi}(l)$ — the probability that k serviced applications will drop at least l applications from the queue outside the safe zone.

$$\tilde{\pi}_1(l) = \sum_{i=l}^{\infty} q(i), \quad \tilde{\pi}_k(l) = \sum_{i=0}^{l-1} \pi_k(i) \tilde{\pi}_1(l-1), \quad k > 1, l \geq 0. \quad (3)$$

a) $0 \leq i \leq Q_1$ — the renovation mechanism is not enabled, leaving the system is only possible by service.

$$p_{i,j} = A_{i+1-j}^{(1)} = \int_0^{\infty} \frac{(\mu x)^{i+1-j}}{(i+1-j)!} e^{-\mu x} dA(x), \quad p_{i,0} = 1 - \sum_{j=1}^{i+1} A_{i+1-j}^{(1)}. \quad (4)$$

b) $i > Q_1$ — the renovation mechanism is enabled at the current moment of arrival. Subcases:

b.1) $Q_1 + 1 < j < i + 1$ — the renovation mechanism will remain enabled by the time of a new application arrival.

$$p_{i,j} = A_{i+1-j}^{(2)} = \int_0^{\infty} \sum_{k=1}^{i+1-j} \pi_k (i+1-j-k) \frac{(\mu x)^k}{k!} e^{-\mu x} dA(x), \quad (5)$$

b.2) $j = Q_1 + 1$ — by the time of a new request arrival, the renovation mechanism will be disabled.

$$p_{i,Q_1+1} = A_{i-Q_1}^{(2)} = \int_0^{\infty} \sum_{k=1}^{i-Q_1} \pi_k (i-Q_1-k) \cdot \frac{(\mu x)^k}{k!} e^{-\mu x} dA(x). \quad (6)$$

b.3) $j = Q_1$ — by the time of a new request arrival, the renovation mechanism will be disabled.

$$\begin{aligned} p_{i,Q_1} = A_{i+1-Q_1}^{(3)} &= \int_0^{\infty} \sum_{k=1}^{i-Q_1} \pi_k (i-Q_1-k) \cdot \frac{(\mu x)^{k+1}}{(k+1)!} e^{-\mu x} dA(x) + \\ &+ \int_0^{\infty} \sum_{k=1}^{i-Q_1} \tilde{\pi}_k (i-Q_1-k) \cdot \frac{(\mu x)^k}{k!} dA(x), j \neq i+1. \end{aligned} \quad (7)$$

b.4) $0 < j < Q_1$ — by the time of a new request arrival, the renovation mechanism will be disabled.

$$p_{i,j} = A_{i+1-j}^{(4)} = \int_0^{\infty} \int_0^x A_{i-1-Q_1}^{(3)}(y) dy A_{Q_1-j}^{(1)}(x-y) dA(x), \quad (8)$$

where

$$\begin{aligned}
 A_{i-1-Q_1}^{(3)}(y) &= \sum_{k=1}^{i-Q_1} \pi_k(i-Q_1-k) \cdot \frac{(\mu y)^{k-1}}{(k-1)!} e^{-\mu y} + \sum_{k=1}^{i-Q_1-1} \tilde{\pi}_k(i-Q_1-k) \cdot \frac{(\mu y)^k}{k!} e^{-\mu y}, \\
 A_{Q_1-j}^{(1)}(x-y) &= \frac{(\mu(x-y))^{Q_1-j}}{(Q_1-j)!} e^{\mu(x-y)}, \\
 p_{i,0} &= 1 - \sum_{j=1}^{i+1} p_{i,j} = 1 - \left(\sum_{j=1}^{Q_1-1} A_{i+1-j}^{(4)} + A_{i+1-Q_1}^{(3)} \sum_{j=Q_1+1}^{i+1} A_{i-Q_1}^{(2)} \right). \quad (9)
 \end{aligned}$$

3. Stationary distribution of the number of applications in the system over the embedded Markov chain

Let $\vec{\pi}$ be the stationary distribution of the number of applications in the system over the embedded Markov chain (just before the receipt of a new application in the system).

System of equilibrium equations in matrix form is

$$\vec{\pi} = \vec{\pi} P$$

and in scalar form is

$$\begin{cases} \pi_0 = \sum_{i=0}^{\infty} A_i^{(5)} \cdot p_{i,i}, \\ \pi_k = \sum_{i=k-1}^{Q_1-1} \pi_i \cdot A_{i+1-k}^{(1)} + \pi_{Q_1} \cdot A_{i+1-Q_1} + \sum_{i=Q_1+1}^{\infty} \pi_i \cdot A_{i+1-Q_1}^{(3)}, \\ \pi_{Q_1} = \pi_{Q_1-1} A_0 + \pi_{Q_1} \cdot A_1^{(3)} + \sum_{i=Q_1}^{\infty} \pi_i A_{i+1-Q_1}^{(3)}, \\ \pi_k = \sum_{i=k-1}^{\infty} \pi_i \cdot A_{i+1-k}^{(2)}, k > Q_1. \end{cases}. \quad (10)$$

As in [9, 10, 13] we assume that π_k can be representable as

$$\pi_k = g^{k-Q_1-1} \pi_{Q_1+1}, \quad (k > Q_1), \quad (11)$$

where g some constant defined by the equation 12.

$$g = \alpha(\mu(1 - g\pi(g))), 0 < g < 1. \quad (12)$$

4. Probabilistic characteristics of the system

Let $p^{(serv)}$ be the probability that the application received in the system will be serviced and let $p^{(loss)}$ be the probability that the application received in the system will be dropped.

$$p^{(serv)} + p^{(loss)} = 1.$$

To calculate them, we introduce auxiliary probabilities $p_i^{(serv)}$ and $p_i^{(loss)}$ — the probability that the application, that found in the system at the time of its arrival i , ($i \geq 0$) other applications, will be served or dropped.

$$p^{(serv)} = \sum_{i=0}^{\infty} p_i^{(serv)} \pi_i, \quad p^{(loss)} = \sum_{i=0}^{\infty} p_i^{(loss)} \pi_i.$$

1. If, at the time of receipt of the application under consideration, the threshold value Q_1 has not been overcome (i.e. the renovation mechanism is not enabled), then the application will enter into the safe zone ($0 \leq i \leq Q_1$), so

$$p_i^{(serv)} = 1, \quad p_i^{(loss)} = 0.$$

2. If $i \geq Q_1 + 1$, then at the time of receipt of the application, the safe zone is completely filled and the incoming application can be dropped in the future.

$$p_i^{(serv)} = \sum_{k=1}^{i-Q_1-1} \sum_{l=0}^{i-Q_1-k} \pi_k(l), \quad p_i^{(loss)} = \sum_{k=1}^{i-Q_1-1} \tilde{\pi}_k(i - Q_1 - k).$$

As result we obtain:

$$p^{(serv)} = 1 - \pi_{Q_1+1} \cdot \frac{1 - \hat{\pi}(g)}{(1-g)(1-g\hat{\pi}(g))}, \quad p^{(loss)} = \pi_{Q_1+1} \cdot \frac{(1 - \hat{\pi}(g))}{(1-g)(1-g\hat{\pi}(g))}. \quad (13)$$

5. Time characteristics

Let $W^{(serv)}(x)$ be the distribution function of waiting time for the start of service by the application received in the system, $W^{(loss)}(x)$ — be the distribution function of the time spent in the queue by the dropped application.

$W_i^{(serv)}(x)$ and $W_i^{(loss)}(x)$ — conditional distribution functions of the time spent in the queue by the application, that found at the moment of its arrival i ($i \geq 1$) other applications in the system, and which will bw servwed or droppe respectively.

$$W^{(serv)}(x) = \frac{1}{p^{(serv)}} \sum_{i=0}^{\infty} W_i^{(serv)}(x) \cdot \pi_i, \quad W^{(loss)}(x) = \frac{1}{p^{(loss)}} \sum_{i=0}^{\infty} W_i^{(loss)}(x) \cdot \pi_i. \quad (14)$$

We also will use the Laplace-Stieltjes transforms $\omega^{(serv)}(s)$, $\omega^{(loss)}(s)$, $\omega_i^{(serv)}(s)$ and $\omega_i^{(loss)}(s)$ of functions $W^{(serv)}(x)$, $W^{(loss)}(x)$, $W_i^{(serv)}(x)$ and $W_i^{(loss)}(x)$ respectively.

5.1. Serviced task. a) $i = 0$ (the system is empty at the time of receipt of the application under consideration)

$$W_0^{(serv)}(x) = 1, x \geq 0.$$

b) $0 < i \leq Q_1$ (the system is not empty, but there is at least one free space in the safe zone, the renovation mechanism is not enabled)

$$W_i^{(serv)}(x) = H_i(x) = \int_0^{\infty} \frac{\mu^i x^{i-1}}{(i-1)!} e^{-\mu x} dx, \quad x \geq 0, i \geq 1.$$

$$\omega_i^{(serv)}(s) = \left(\frac{\mu}{\mu + s} \right)^i, \quad i \geq 1.$$

c) $i \geq Q_1 + 1$ (at the time of receipt of the application under consideration, the safe zone is filled and the renovation mechanism is enabled).

$$W_{Q_1+i}^{(serv)}(x) = \sum_{j=0}^i H_{Q_1+j}(x) \cdot \pi_j(i-j),$$

$$\omega_{Q_1+i}^{(serv)}(s) = \sum_{j=0}^i \left(\frac{\mu}{\mu + s} \right)^{Q_1+i} \cdot \pi_j(i-j).$$

Then

$$\begin{aligned} \omega^{(serv)}(s) &= \frac{1}{p^{(serv)}} \sum_{i=0}^{\infty} \omega_i^{(serv)}(s) \cdot \pi_i = \\ &= \frac{1}{p^{(serv)}} \left(\pi_0 + \sum_{i=0}^{Q_1} \cdot \left(\frac{\mu}{\mu + s} \right)^i + \pi_{Q_1+1} \left(\frac{\mu}{\mu + s} \right)^{Q_1+1} \cdot \frac{\hat{\pi}(g)(\mu + s)}{\mu + s - \mu g \hat{\pi}(g)} \right) \end{aligned} \quad (15)$$

and the average waiting time for the start of service is:

$$W^{(serv)} = \frac{1}{p^{(serv)}} \left(\frac{1}{\mu} \sum_{i=1}^{Q_1} i \pi_i + \frac{\pi_{Q_1+1} \cdot \hat{\pi}(g) \cdot (Q_1 + 1 - Q_1 g \hat{\pi}(g))}{\mu(1 - g \hat{\pi}(g))^2} \right). \quad (16)$$

5.2. Time characteristics for a dropped application. If the system is empty or there is at least one place in the safe zone in the queue, then the application entering the system cannot be dropped:

$$W_i^{(loss)}(x) = 0, \quad 0 \leq i \leq Q_1.$$

If the safe zone is completely filled (so the general renovation mechanism is enabled) at the time of arrival of the considered application, then and only then the application under consideration may be dropped:

$$W_{Q_1+1}^{(loss)}(x) = \tilde{\pi}_1(1) \cdot H_1(x),$$

$$\begin{aligned} W_i^{(loss)}(x) &= \tilde{\pi}_1 \cdot (i - Q_1) \cdot H_1(x) + \\ &+ \sum_{k=2}^{i-Q_1} H_k(x) \cdot \sum_{j=0}^{i-Q_1-k} \pi_{k-1}(j) \cdot \tilde{\pi}_1(i - Q_1 + 1 - k - j) \quad , i > Q_1 + 1. \end{aligned}$$

In terms of the Laplace-Stieltjes transform we get:

$$\omega_i^{(loss)}(s) = 0, \quad 0 \leq i \leq Q_1, \quad \omega_{Q_1+1}^{(loss)}(s) = \tilde{\pi}_1(1) \frac{\mu}{\mu + s},$$

$$\begin{aligned} \omega_i^{(loss)}(s) &= \tilde{\pi}_1(i - Q_1) \cdot \frac{\mu}{\mu + s} + \\ &+ \sum_{k=2}^{i-Q_1} \left(\frac{\mu}{\mu + s} \right)^k \sum_{j=0}^{i-Q_1-k} \pi_{k-1}(j) \cdot \tilde{\pi}_1(i - Q_1 + 1 - k - j), \quad i > Q_1 + 1. \end{aligned}$$

The final Laplace-Stieltjes transform for the distribution function of the time spent in the queue by the dropped application:

$$\omega^{(loss)}(s) = \frac{1}{p^{(loss)}} \sum_{i=0}^{\infty} \omega_i^{(loss)}(s) \pi_i = \frac{\pi_{Q_1+1}}{p^{(loss)}} \frac{\mu}{\mu + s} \cdot \frac{1 - \hat{\pi}(g)}{1 - g} \cdot \frac{\mu + s}{\mu + s - \mu g \hat{\pi}(g)}. \quad (17)$$

The average time spent in the queue by a dropped application

$$W^{(loss)} = - \left(\omega^{(loss)}(s) \right)'_{s=0} = \frac{\pi_{Q_1+1}}{p^{(loss)}} \cdot \frac{1 - \hat{\pi}(g)}{1 - g} \cdot \frac{1}{\mu(1 - g \hat{\pi}(g))^2}. \quad (18)$$

5.3. Time characteristics of an arbitrary application. Let $W(x)$ be the distribution function of the time spent in the queue by an arbitrary application.

Then:

$$W(x) = p^{(serv)} \cdot W^{(serv)}(x) + p^{(loss)} \cdot W^{(loss)}(x). \quad (19)$$

In terms of Laplace-Stieltjes transform

$$\begin{aligned} \omega(s) &= p^{(serv)} \cdot \omega^{(serv)}(s) + p^{(loss)} \cdot \omega^{(loss)}(s) = \sum_{i=0}^{Q_1} \pi_i \left(\frac{\mu}{\mu+s} \right)^i + \\ &+ \pi_{Q_1+1} \cdot \frac{\mu+s}{\mu+s - \mu g \hat{\pi}(g)} \left(\left(\frac{\mu}{\mu+s} \right)^{Q_1+1} \hat{\pi}(g) + \frac{\mu}{\mu+s} \cdot \frac{1-\hat{\pi}(g)}{1-g} \right). \end{aligned} \quad (20)$$

Let w be the average time spent in the queue by an arbitrary application.

$$w = p^{(serv)} \cdot w^{(serv)} + p^{(loss)} \cdot w^{(loss)}, \quad (21)$$

$$w = \frac{1}{\mu} \sum_{i=1}^{Q_1} i \pi_i + \frac{1}{\mu} \pi_{Q_1+1} \cdot \frac{1 + Q_1 \hat{\pi}(g)(1-g)}{(1-g)(1-g \hat{\pi}(g))}. \quad (22)$$

6. Conclusion

In this paper, we considered a single-line queuing system with an infinite capacity queue, with a threshold and renovation mechanism. Analytical expressions of the distribution of the number of applications in the system were found for this system, expressions for calculating time and probabilistic characteristics were obtained.

The results obtained for the considered system coincide with the results presented in the paper [10], if threshold value $Q_1 = 0$.

In the future, it is planned to study the time characteristics of the system for the following service and renovation options:

- service in the order of arrival (first come, first serve basis), applications are dropped in reverse order (starting from the last one);
- service in reverse order (last come, first serve basis), applications are dropped in direct order (from the first one);
- renovation and service in reverse order.

REFERENCES

1. Floyd S., Jacobson V. Random Early Detection Gateways for Congestion Avoidance // IEEE/ACM Transactions on Networking. 1993. V. 4 (1). P. 397–413.

2. Ramakrishnan K., Floyd S., Black D. The Addition of Explicit Congestion Notification (ECN) to IP. RFC 3168. Internet Engineering Task Force. 2001. <https://tools.ietf.org/html/rfc3168>
3. Floyd S., Gummadi R., Shenker S. Adaptive RED: An Algorithm for Increasing the Robustness of RED's Active Queue Management. 2001. <http://www.icir.org/floyd/papers/adaptiveRed.pdf>
4. Baker F., Fairhurst G. IETF Recommendations Regarding Active Queue Management. RFC 7567. Internet Engineering Task Force. <https://tools.ietf.org/html/rfc7567>.
5. Adams R. Active Queue Management: A Survey. // IEEE Communications Surveys & Tutorials. 2013. V. 15 (3). P. 1425–1476.
6. Menth M., Veith S. Active Queue Management Based on Congestion Policing (CP-AQM) // In: Measurement, Modelling and Evaluation of Computing Systems. MMB 2018. Lecture Notes in Computer Science. 2018. V. 10740. P. 173–187.
7. Konovalov M. G., Razumchik R. V. Numerical Analysis of Improved Access Restriction Algorithms in a $GI|G|1|N$ System // Journal of Communications Technology and Electronics. 2018. V. 63 (6). P. 616–625.
8. Kreinin A. Queueing Systems with Renovation // Journal of Applied Math. Stochast. Analysis. 1997. V. 10 (4). P. 431–443.
9. Bocharov P. P., Zaryadov I. S. Probability Distribution in Queueing Systems with Renovation // Bulletin of Peoples' Friendship University of Russia. Series "Mathematics. Information Sciences. Physics". 2007. No. 1-2. P. 15–25.
10. Zaryadov I. S., Pechinkin A. V. Stationary Time Characteristics of the $GI/M/n/\infty$ System with Some Variants of the Generalized Renovation Discipline // Automation and Remote Control. 2009. V. 70 (12). P. 2085–2097.
11. Konovalov M., Razumchik R. Comparison of two active queue management schemes through the $M/D/1/N$ queue // Informatika i ee Primeneniya (Informatics and Applications). 2018. V. 12 (4). P. 9–15.
12. Hilquias Viana C. C., Zaryadov I. S., Milovanova T. A., Tsurlukov V. V., Korolkova A. V., Kulyabov D. S.. The General Renovation as the Active Queue Management Mechanism. Some Aspects and Results // In: Communications in Computer and Information Science. 2019. V. 1141. P. 488–502. doi:10.1007/978-3-030-36625-4_39.
13. Hilquias Viana C. C., Zaryadov I. S., Milovanova T. A. Two Types of Single-Server Queueing Systems with Threshold-Based Renovation Mechanism // In: Lecture Notes in Computer Science. 2021. V. 13144. P. 96–210.
14. Gorbunova A., Lebedev A. Queueing System with Two Input Flows, Preemptive Priority, and Stochastic Dropping// Autom Remote Control. 2020. V. 81. P. 2230–2243.doi:10.1134/S0005117920120073

UDC: 510.644;51-74

Feature Selection for a Fuzzy Classification Model Based on a Genetic Algorithm

O.N. Kochueva¹

¹National University of Oil and Gas "Gubkin University", 65 Leninsky Prospekt,
Moscow, 119991, Russia

kochueva.o@gubkin.ru

Abstract

The paper presents a new method of feature selection for a classification model based on a symbolic regression method and genetic algorithm. For complex practical problems, building a unified predictive model for various states of a system or a process encounters a lot of difficulties, but the task can be divided into 2 stages: a)to obtain a classification of system states; b)to build models with good predictive qualities for each class. The fuzzy approach makes it possible to specify states (set of parameters) which can be assigned to more than one class. A novelty of the presented model is the use of symbolic regression to identify a set of input variables for the classification model. The algorithm is accompanied by an example of practical application.

Keywords: machine learning; feature selection; symbolic regression; fuzzy classification model; knowledge extraction; genetic algorithm

1. Introduction and literature review

Classification problem is one of the main components of knowledge extraction. Nowadays, with the development of measurement technology, a large amount of data is available, and it is necessary to use a special strategy for specifying parameters or their combinations as the feature variables of the model. In empirical classification problems, the number of the feature variables is unknown, and the choice of input variables for classification poses its own challenges especially when there is an effect of reciprocal influence of several parameters.

Fuzzy classification model can be a powerful instrument and provide an insight for practical problems when crisp classification reaches its limits. One example of such a situation is the classification of emissions of harmful pollutants, where a crisp approach can define a certain amount of a pollutant as safe, while its increase by 0.01% can exceed the limit and be classified as dangerous. The use of fuzzy logic

makes it easy to overcome this disadvantage - the set of initial data can be assigned to two classes at once, for each of them, the degree of membership can be determined.

The application of classification problem is extremely wide and fuzzy approach has given good results in a number of real world applications (examples can be found in [1], [2], [3]). An overview of algorithms for fuzzy classification models is given in [4]. In most of the algorithms proposed by the authors (see [5], for example) two stages can be considered - feature extraction step and inference step. Optimization algorithms are used to determine the distribution of fuzzy sets for each feature variable at the first step. At the second step, the optimization procedure helps determine weights for each rule to obtain the best classification result.

The popular optimization algorithms for multivariate functions applied to the solution of the classification problem: particle swarm optimization technique [1]; grey wolf optimization algorithm [2]; artificial bee colony [6]; various modifications of Genetic Algorithm (GA) [5], [7]. The main ideas to use genetic programming (GP) methodology in fuzzy inference system (FIS) design are presented in [8].

The key issue in building a classification model is to determine the set of input variables. This paper presents a new approach to extracting feature variables for a classification model from formulas obtained using the symbolic regression (SR) method.

2. Methodology

The SR is one of the machine learning methods. It allows generate models in the form of analytic equations, while it is not required to predetermine the structure of the model, and the interaction between variables is transparent. The idea of SR is based on a genetic algorithm [9], and a model is built as a sequence of chromosomes. Chromosomes are generated randomly from a set of genes, as a gene can be used a predictor, a number, an arithmetic operation or a function. At the first step a population of functions $F_l, l = 1, \dots, N$ (N – the size of population) is randomly generated, then for each individual F_l the fitness function FF_l is calculated as follows

$$FF_l = - \sum_{i=1}^n (F_l(x_{1,i}, x_{2,i}, \dots, x_{m,i}) - y_i)^2, \quad (1)$$

where $x = \{x_1, x_2, \dots, x_m\}$ – a set of input variables, y – a dependent variable, n – number of training samples. The sum of squared differences between the observed dependent variable y_i and the value predicted by function F_l is taken with a negative sign, since a larger value of the fitness function is favorable for the individual. The function FF_l determines the viability of a particular individual F_l in the population,

which means the ability of the function F_l to correctly predict the value of the target variable y . Then, from the functions F_l , the parents are selected, and new chromosomes are created with the crossover operation. The next step of GA is mutation (transformation of a chromosome that accidentally changes one or more of its genes), this operation avoids falling to a local extremum. The fitness function (1) for new individuals is calculated and the next generation is formed. The described procedure is repeated until the change in the best value of the fitness function becomes less than a given tolerance or a predetermined number of generations is obtained.

So using the SR methodology, a predictive model in the form of an analytic equation can be built, but often it is not possible to obtain an acceptable value of the mean absolute error (MAE) for the entire range of input parameters. The way out is to allocate separate classes for the system or the process under study, after which its own model can be built for each class, and the fuzzy classification model is necessary to determine for the input data the degrees of membership to each class.

Let k – the number of classes, μ_i – the degree of membership to the i -th class (the output of the fuzzy classification model), then

$$y = \sum_{i=1}^k F_i(x_1, x_2, \dots, x_m) \cdot \mu_i. \quad (2)$$

Consider the stages of building a fuzzy classification model and describe the procedure of feature selection:

- 1) Determine the required number of classes k . Mark up the data set according to the selected number of classes.
- 2) For each class i , using the SR method, obtain a formula in the form

$$F_i(x_1, \dots, x_m) = f_{i,1}(x_1, \dots, x_m) + f_{i,2}(x_1, \dots, x_m) + \dots + f_{i,m}(x_1, \dots, x_m) \quad (3)$$

A special case of the function $f_{i,j}$ can be a function of one input variable, for example $f_{i,j}(x_1, \dots, x_m) = C \cdot x_m$.

- 3) Analyze the chromosome-functions $f_{i,j}$ and select those that will be input variables for classification, since they have the greatest differences for different classes. The further procedure will be described for the case of two classes, since the practical problem that initiated this study required the consideration of just two classes.

For the samples of the training set belonging to the first class, calculate the set of the values $f_{i,j}^{(1)}(x_1^{(1)}, \dots, x_m^{(1)})$ and for the second class $f_{i,j}^{(2)}(x_1^{(2)}, \dots, x_m^{(2)})$. If the value sets $f_{i,j}^{(1)}$ and $f_{i,j}^{(2)}$ do not intersect for at least one pair (i, j) , the function

$f_{i,j}$ is a feature variable for a crisp classification. In the case of intersection of sets, it is necessary to build a fuzzy classification model and choose which functions $f_{i,j}$ will be input variables. Therefore, we need a numerical criterion to determine the quality of division into classes, if $f_{i,j}$ is a feature variable. Let a median of values $f_{i,j}^{(1)}$ for training set for the first class is greater than a median of values $f_{i,j}^{(2)}$ for the second class, let $s_{i,j} \in [\min(f_{i,j}^{(1)}), \max(f_{i,j}^{(2)})]$, $K_1(s_{i,j})$ – the number of training set samples belonging to the first class for which $f_{i,j}^{(1)} < s_{i,j}$, $K_2(s_{i,j})$ – the number of training set samples belonging to the second class for which $f_{i,j}^{(2)} > s_{i,j}$, n_i – the number of samples in i -th class. The function

$$g_{i,j}(s_{i,j}) = K_1(s_{i,j})/n_1 + K_2(s_{i,j})/n_2 \rightarrow \min \quad (4)$$

tends to 0 if there is a clear difference between the sets of values $f_{i,j}^{(1)}$ and $f_{i,j}^{(2)}$. Thus, for two classes, the problem turns to finding the minimum of a function of one variable on a fixed interval, the algorithm based on golden section search can be used. The value $g_{i,j}$ determines to what extent the chromosome $f_{i,j}$ is suitable for the set of FIS input variables.

- 4) Determine the parameters of membership functions and form a system of rules for a FIS. For two classes, two terms ("LOW" and "HIGH") are constructed for each input variable as shown in Fig 1.

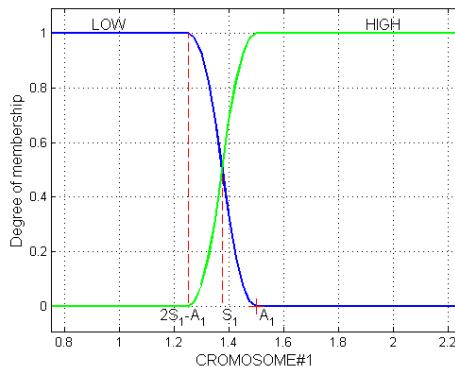


Fig. 1. Membership functions for terms "LOW" and "HIGH" and their parameters.

Thus, the point $s_{i,j}$ is determined when searching for the minimum of function (4), and the membership functions are assumed to be symmetrical, then it remains for each input variable of the FIS to determine the parameter A_j .

The rules for the FIS are formulated according to the "IF-THEN" scheme. An example of three rules is shown below:

IF CHROMOSOME#1 is LOW THEN CLASS is FIRST;

IF CHROMOSOME#2 is LOW THEN CLASS is SECOND;

IF CHROMOSOME#3 is HIGH THEN CLASS is FIRST;

Each rule is assigned a weight w_j corresponding to its contribution to the correct decision.

The effectiveness of a classification model is determined by metrics such as *Precision*, *Recall*, and F_1 -score, that can be calculated as follows:

$$Precision = \frac{T1}{T1 + F1}, Recall = \frac{T1}{T1 + F2}, F_1 = 2 \frac{Precision \cdot Recall}{Precision + Recall}, \quad (5)$$

where $T1$ —the number of first class predictions that actually belong to the first class; $F1$ —the number of first class predictions that actually belong to the second class; $F2$ —the number of second class predictions that actually belong to the first class.

The parameters of membership functions A_j and the weights w_j are determined from the solution of the constrained optimization problem. The goal is to maximize the performance of the fuzzy classification model, namely the value of F_1 -score (5), with the control parameters $A_j, w_j, (j = \overline{1, J})$ for each input variable (chromosome-function) and corresponding rule. The constraints are:

$$\begin{aligned} s_{i,j} \leq A_j \leq max(f_{i,j}(x_1, x_2, \dots, x_m)), \\ 0 \leq w_j \leq 1. \end{aligned} \quad (6)$$

To solve the multidimensional optimization problem, a GA can be used. Thus, the optimal set of FIS parameters is obtained, it gives the highest value for the F_1 -score (5) for the training set, and then, using fuzzy inference, for any set of input data (x_1, x_2, \dots, x_m) , the degree of membership to the first and second classes can be calculated.

- 5) To improve the quality metrics of the resulting model, new chromosome function $f_{i,j}$ should be added to the set of FIS input variables, and it is necessary to determine the membership functions parameters for the new feature variable, modify the system of rules for FIS, and calculate the quality metrics. The choice of new functions should be carried out in accordance with the values of function (4).

3. Example and discussion

The described model was tested in the classification of carbon monoxide emissions from gas turbines, the data for the study is presented in the open data repository [10]. The dataset was collected for 5 years, it contains 36,733 instances of 11

sensor measures aggregated over one hour, including three external environmental parameters (air temperature (AT), air humidity (AH), atmosphere pressure (AP)), six indicators of the gas turbine technological process (air filter difference pressure (AFDP), gas turbine exhaust pressure (GTEP), turbine inlet temperature (TIT), turbine after temperature (TAT), turbine energy yield (TEY), compressor discharge pressure (CDP)), and two target variables (emissions of carbon monoxide (CO) and the total emissions of nitrogen monoxide and nitrogen dioxide (NOx)). Description of the dataset is given in [11], some new details are given in [12].

CO emissions can increase dramatically up to 20–40 times the median value, so it is important to indicate the situation that leads to extraordinary emissions. So the dataset was split into 2 classes: "Standard" - for emissions less than 5 mg/m^3 (about 90% of data) and "Extreme" for the higher values. For each class, a separate predictive model was built using the SR methodology. To build a model the input variables were standardized using the Z-score. The formula obtained for the data set for the "Extreme" class is

$$F_{extr} = 225.59 \cdot GTEP^2 \cdot TAT - 20.67 \cdot TAT \cdot \exp(-TIT) - 20.67 \cdot AFDP \times TIT + 7.88 \cdot GTEP \cdot \exp(-TIT^2) - 67.05 \cdot AT \cdot AFDP \cdot TIT - 24.68 \cdot AT \times TIT \cdot \exp(-TAT) - 68.81 \cdot AFDP \cdot GTEP \cdot TIT \cdot \exp(-GTEP) + 7.25. \quad (7)$$

Histograms demonstrating the distribution of the values of the chromosome-functions $f_{1,1}$ and $f_{1,3}$ (CHROMOSOME#1 and CHROMOSOME#3) are shown in Fig. 2. The red dashed lines correspond to the values $s_{1,1}$ and $s_{1,3}$ determined when optimizing the function (4).

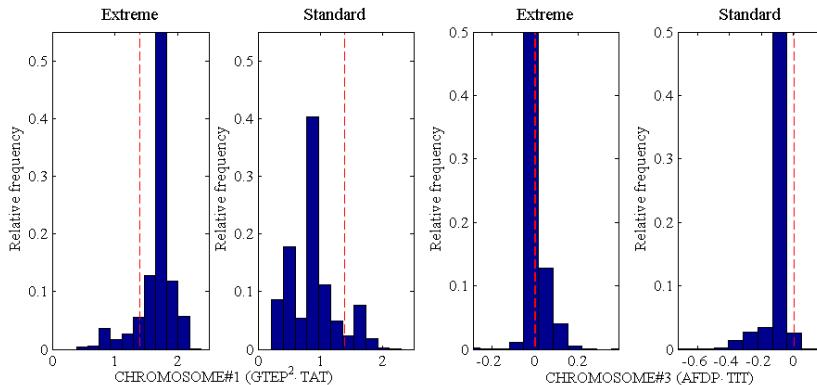


Fig. 2. Histograms of chromosomes #1 and #3

For CHROMOSOME#1, to the left of the point $s_{1,1}$ there are 87% of the values of $GTEP^2 \cdot TAT$ for the data belonging to the "Standard" class for the training set, and to the right of this point, there are more than 90% of the values of $GTEP^2 \cdot TAT$ for the data belonging to the "Extreme" class. So CHROMOSOME#1 was chosen as input variables for FIS. For CHROMOSOME#3, to the left of the point $s_{1,3}$ there are 95% of the values $AFDP \cdot TIT$ for data belonging to the "Standard" class for the training set, and to the right of this point, 46% of the values $AFDP \cdot TIT$ for data belonging to the "Extreme" class are located. Thus, CHROMOSOME#3 does not give a noticeable difference between the classes, as can be seen in Fig. 2 and when comparing the values of $g_{i,j}(s_{i,j})$: $g_{1,1}(s_{1,1}) = 0.223$, $g_{1,3}(s_{1,3}) = 0.583$. Therefore, CHROMOSOME#3 cannot be selected as an input variable for FIS. In addition to CHROMOSOME#1, two more chromosomes - functions were included in the set of FIS input variables.

For the test set, F_1 -score for the "Extreme" class is 0.82, which exceeds the value obtained using the Random Forest (RF) classification algorithm for the same dataset. The fuzzy classification model, combined with the SR prediction models for each class, can reduce MAE by 10 times, compared to using a single model [13].

4. Conclusion

The paper presents a new feature selection approach for a fuzzy classification model based on a symbolic regression method, fuzzy inference system and genetic algorithm. The advantages of the presented approach are that the formulas obtained using symbolic regression a) can be used as predictive models; b) form a set of feature variables for the fuzzy classification model; c) clarify the interaction of input parameters. The genetic algorithm is used in the framework of symbolic regression and as a tool for solving the problem of multivariate optimization to determine the parameters of the FIS. The model has been applied to real data and the results are superior to the models presented in [11].

REFERENCES

1. Marimuthu S., Mohamed Mansoor Roomi S. Particle Swarm Optimized Fuzzy Model for the Classification of Banana Ripeness // IEEE Sensors Journal. 2017. V. 17 (15). P. 4903–4915.
2. Zou Q, Liao L, Ding Y, Qin H. Flood Classification Based on a Fuzzy Clustering Iteration Model with Combined Weight and an Immune Grey Wolf Optimizer Algorithm //Water. 2019. V. 11(1). 80.
3. Saini, J., Dutta, M., Marques, G. ADFIST: Adaptive Dynamic Fuzzy Inference System Tree Driven by Optimized Knowledge Base for Indoor Air Quality Assessment. // Sensors. 2022. V. 22. 1008. <https://doi.org/10.3390/s22031008>

4. Ducange P., Fazzolari M., Marcelloni F. An overview of recent distributed algorithms for learning fuzzy models in Big Data classification //Journal of Big Data. 2020. V.7(1). 19.
5. Guo N.R., Li T.-H.S. Construction of a neuron-fuzzy classification model based on feature-extraction approach // Expert Systems with Applications. 2011. V. 38 (1). P. 682–691.
6. Feng T.-C., Chiang T.-Y., Li T.-H.S. Enhanced hierarchical fuzzy model using evolutionary GA with modified ABC algorithm for classification problem //ICCSS 2015 - Proceedings: 2015 International Conference on Informative and Cybernetics for Computational Social Systems. 2015. 7281146, P. 40–44.
7. Feng T.-C., Li T.-H.S., Kuo P.-H. Variable coded hierarchical fuzzy classification model using DNA coding and evolutionary programming // Applied Mathematical Modelling. 2015. V. 39 (23-24). P. 7401–7419.
8. Ojha, V.; Abraham, A.; Snasel, V. Heuristic Design of Fuzzy Inference Systems: A Review of Three Decades of Research. // Eng. Appl. Artif. Intell. 2019. V. 85. P. 845—864.
9. Mitchell M. An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA, USA, 1996.
10. Dua D., Graff C. UCI Machine Learning Repository. Available online: <http://archive.ics.uci.edu/ml> (accessed on 10 May 2022).
11. Kaya, H.; Tüfekci, P.; Uzun, E. Predicting CO and NOx emissions from gas turbines: Novel data and a benchmark PEMS // Turk. J. Electr. Eng. Comput. Sci. 2019. V. 27. P. 4783–4796.
12. Kochueva O., Nikolskii K. Data Analysis and Symbolic Regression Models for Predicting CO and NOx Emissions from Gas Turbines //Computation. 2021. V. 9. 139. P. 1–16
13. Kochueva O. Razrabotka modelej prognozirovaniya vybrosov oksidov ugleroda i azota gazovyh turbin na osnove geneticheskikh algoritmov //Delovoj zhurnal Neftegaz.RU. 2022. V. 5-6 (125-126). P. 14–20. (in Russian)

УДК: 519.872

Математическое моделирование передачи данных в сети FANET в виде RQ-систем

Д.А. Плаксин¹, Е.А. Фёдорова¹, О.Д. Лизюра¹, Д.В. Шашев¹, С.П.
Моисеева¹

¹Томский государственный университет, проспект Ленина, 36, Томск, Россия
daniel.plaksin@gmail.com, moiskate@mail.ru, oliztsu@mail.ru, dshashev@mail.ru,
smoiseeva@mail.ru

Аннотация

В работе предлагается математическая модель передачи данных в сетях FANET в виде RQ-системы (Retrial Queue). Представлено описание простейшей RQ-системы с конфликтами заявок. Перечислены основные характеристики и направления усложнения модели, необходимые для анализа реальных сетей связи.

Ключевые слова: FANET, случайный множественный доступ, теория массового обслуживания, RQ-системы

1. Введение

Беспилотные летательные аппараты (БПЛА или дроны) широко используются для решения военных и гражданских задач, таких как поиск людей, видеомониторинг местности при пожарах, поиск разрыва линий электропередач, газопровода, орошение полей,брос с удобрений, метеорологических целей (замер физических показателей на высоте давления, температуры, влажности, контроль токсичных веществ), развертывания "летающих сетей" беспроводной связи и др. По видам связи различают связь между БПЛА и управляющим устройством, дрон-дрон и дрон-спутник. По протоколам передачи данных используются ZigBee, 6LoWPAN, bluetooth, Wi-Fi, LTE, 3G, 4G и др.

Большинство статей, посвященных протоколам доступа в сетях FANET, используют алгоритмические и имитационные модели для определения основных показателей качества обслуживания - Quality of Service (QoS) [1, 2, 3]. Такие модели полезны для сетей со слабо меняющимися параметрами. В ином случае для каждого набора параметров необходимо проводить новую серию расчетов. Применение стохастических моделей позволяет не только рассчитать показатели QoS для конкретных наборов входных параметров сети, но и исследовать

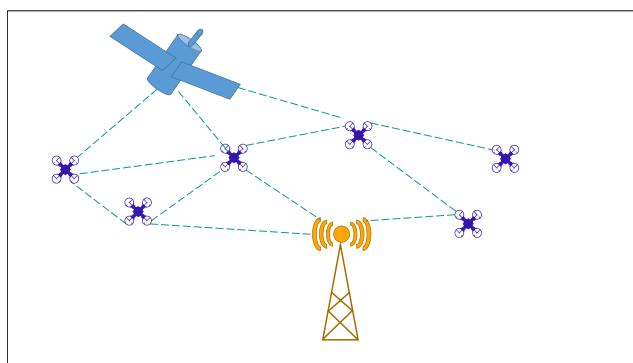


Рис. 1. Сеть FANET

закономерности их изменения. Однако работ, связанных с применением теории массового обслуживания (ТМО) для исследования сетей FANET немного.

В статье [4] для исследования мультитрафика «воздух-земля», обеспечиваемый беспилотными летательными аппаратами, предлагается однолинейная модель ТМО с неограниченным числом мест для ожидания в очереди, но с групповым поступлением заявок и приоритетами. Также моделирование сетей БПЛА в виде систем массового обслуживания рассматривалось в работе [5]. В статье предложена модель в виде однолинейной СМО с очередью с применением методов стохастической геометрии. Получены основные вероятностные характеристики модели, позволяющие оценить величину потерь в сети. А в работе [6] группу БПЛА предлагается рассматривать как сеть Джексона.

Однако передача данных в реальных сетях обычно происходит согласно протоколу множественного случайного доступа [7]. То есть применение неклассических моделей теории массового обслуживания - RQ-систем (retrying queueing system) - должно дать лучший результат.

Стоит отметить, что в литературе были некоторые попытки учесть эффект множественного доступа в моделях. Так в статье [8] предлагается пороговый протокол множественного доступа (channel threshold based multiple access, CTMA), который учитывает наличие нескольких уровней приоритета пакетов и моделируется в виде однолинейной СМО $M/G/1$ с классификатором трафика, несколькими буферами и прогулками прибора. Такая модель также не учитывает особенности протоколов случайного множественного доступа, однако, как показывают авторы, позволяет избежать коллизий между заявками разного приоритета.

Работа [9] также посвящена исследованию протоколов случайного множественного доступа в сетях FANET. Узел сети моделируется с помощью однолинейной СМО с простейшим входящим потоком, детерминированной длительностью об-

служивания и конечной очередью. Здесь также не учитываются особенности мобильных сетей связи и возможность возникновения коллизий.

В данной статье мы предлагаем моделировать узел сети FANET в виде RQ-систем. RQ-системы являются относительно новыми моделями теории массового обслуживания, в которых нет очереди в классическом понимании. Вместо нее есть некоторое виртуальное место, называемое орбитой, в которой реализуется множественный доступ (обычно случайного характера). То есть каждая заявка в любой момент времени имеет возможность обратиться к обслуживающему устройству.

2. Математическая модель

Рассмотрим модель подробнее на примере простейшей RQ-системы с конфликтами (Рис. 2). Пусть на вход системы поступает некоторый поток заявок (то есть сигналов/ пакетов данных от БПЛА или управляющего устройства). Система имеет один прибор (это канал связи, по которому передаются данные). Если прибор свободен в момент поступления заявки, то заявка занимает его для обслуживания (то есть осуществляется передача данных от источника к приемнику). Если канал связи был занят, то возникает конфликт (коллизия) и обе заявки (поступившая и обслуживающаяся) уходят на орбиту, где осуществляют некоторую случайную задержку, после которой вновь пытаются передать данные. Любая заявка на орбите (то есть любой объект сети) имеет возможность в любой момент времени обратиться к прибору (таким образом, имеет место быть случайный множественный доступ).

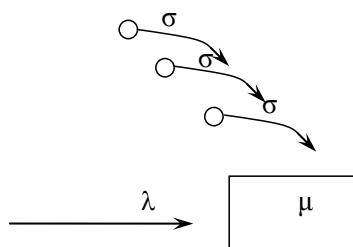


Рис. 2. Математическая модель

В самой простейшей модели предполагается, что законы распределения случайных величин обслуживания, задержки на орбите экспоненциальны с пара-

метрами μ и σ , а сам входящий поток заявок имеет распределение Пуассона с параметром λ . Однако модель может быть усложнена, в зависимости от реальных данных, например, непуассоновские потоки и /или неэкспоненциальное распределение времени задержки и обработки данных. Кроме того, возможно учесть следующие моменты, возникающие в сетях FANET:

- Случайный размер передаваемых данных (в терминах ТМО это неординарные потоки требований).
- Разнородность данных (несколько входящих потоков).
- Приоритетность в обслуживании (приоритетный поток заявок с вытеснением).
- Потеря данных по причине их неактуальности (нетерпеливые заявки) и др.

Аппарат теории массового обслуживания позволяет вывести аналитические формулы для основных показателей качества функционирования сети, в качестве которых выбраны следующие:

- вероятность потерь пакетов и задержки доставки данных,
- среднее время передачи данных (время пребывания в системе),
- доля повторных вызовов на одну первичную,
- пропускная способность сети и др.

3. Заключение

Таким образом, в работе предлагается использовать системы массового обслуживания с повторными вызовами (RQ-системы) в качестве математических моделей сетей FANET. В работе представлено описание простейшей RQ-системы с конфликтами заявок, а также перечислены направления усложнения модели в случае анализа реальных сетей связи.

Исследование выполнено при поддержке Программы развития Томского государственного университета (Приоритет-2030).

Литература

1. K. A. Darabkh, M. G. Alfawares, S. Althunibat, Mdrma: Multi-data rate mobility-aware aodv-based protocol for flying ad-hoc networks, *Vehicular Communications* 18 (2019) 100163.
2. A. V. Leonov, V. O. Ryabchevsky, Performance evaluation of aodv and olsr routing protocols in relaying networks in organization in mini-uavs based fanet: Simulation-based study, in: 2018 Dynamics of Systems, Mechanisms and Machines (Dynamics), IEEE, 2018, pp. 1–6.

3. A. V. Leonov, Application of bee colony algorithm for fanet routing, in: 2016 17th International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices (EDM), IEEE, 2016, pp. 124–132.
4. Z. Zhang, D. Wu, W. Xu, J. Shang, Z. Feng, P. Zhang, Uav-enabled multiple traffic backhaul based on multiple rans: A batch-arrival-queuing-inspired approach, *IEEE Access* 7 (2019) 161437–161448.
5. Р. Киричек, А. Парамонов, Беспилотный летательный аппарат как система массового обслуживания, *Электросвязь* (7) (2015) 16–19.
6. R. Kirichek, A. Paramonov, A. Koucheryavy, Swarm of public unmanned aerial vehicles as a queuing network, in: V. Vishnevsky, D. Kozyrev (Eds.), *Distributed Computer and Communication Networks*, Springer International Publishing, Cham, 2016, pp. 111–120.
7. В. В. Бородин, А. М. Петраков, В. А. Шевцов, Анализ эффективности передачи данных в сети связи группировки беспилотных летательных аппаратов, *Труды МАИ* (81) (2015) 27–46.
8. B. Zheng, K. Zhuo, H. Wu, T. Xie, Multi-priority queueing mechanism for channel threshold based multiple access in fanets, in: 2021 IEEE 9th International Conference on Information, Communication and Networks (ICICN), IEEE, 2021, pp. 15–20.
9. T. Jinhui, W. Yequn, D. Shufu, S. Qilu, G. Huang, A feedback-retransmission based asynchronous frequency hopping mac protocol for military aeronautical ad hoc networks, *Chinese Journal of Aeronautics* 31 (5) (2018) 1130–1140.

УДК: 519.872

Асимптотически-диффузионный анализ RQ-системы ММРР/М/1 с разнотипными вызываемыми заявками

А.А. Назаров¹, Т. Фунг-Дук², С.В. Пауль¹, О.Д. Лизюра¹

¹Национальный исследовательский Томский государственный университет,
проспект Ленина, 36, Томск, Россия

²Университет Цукуба, Цукуба, Япония

nazarov.tsu@gmail.com, tuan@sk.tsukuba.ac.jp, paulsv82@mail.ru, oliztsu@mail.ru

Аннотация

В статье рассматривается RQ-система с разнотипными вызываемыми заявками и одним обслуживающим прибором. В систему поступает ММРР-поток заявок. Заявка входящего потока занимает прибор для обслуживания, если он свободен. В ином случае, поступившая заявка присоединяется к виртуальному месту для ожидания, называемому орбитой. Повторные обращения с орбиты происходят независимо друг от друга после случайной задержки. Когда прибор свободен, он также может вызывать заявки сам. Вызываемые заявки вызываются и обслуживаются с интенсивностью, зависящей от их типа. В статье представлено исследование числа заявок на орбите в описанной системе с помощью метода асимптотически-диффузионного анализа.

Ключевые слова: RQ-система, метод асимптотически-диффузионного анализа, диффузионная аппроксимация, вызываемые заявки

1. Введение

RQ-системы (retrial queues, модели массового обслуживания с повторными обращениями) с вызываемыми заявками являются математической моделью смешанного call-центра. В сравнении с традиционными call-центрами, где операторы принимают звонки от клиентов, в смешанном call-центре оператор может совершать звонки для проведения опросов, рекламных предложений и т. д. Такой подход позволяет уменьшить время простоя в периоды малой загрузки центра. Выбор именно RQ-систем в качестве модели не случаен, так как повторные обращения позволяют учесть особенности поведения заявок при обслуживании по телефону. Статья [1] посвящена тому, как повторные обращения влияют на моделирование обслуживающих систем.

Ранее RQ-системы с вызываемыми заявками рассматривались в предположении замкнутости системы, то есть когда имеется ограниченное количество внешних источников поступающих звонков. Имитационному моделированию таких систем посвящены работы [2, 3, 4].

Модели с вызываемыми заявками также рассматривались в предположении, что прибор вызывает заявки не из сторонних источников, а из орбиты [5, 6].

Также вызывают интерес RQ-системы с конечной орбитой. Исследование такой системы с произвольным распределением времени обслуживания представлено в статье [7].

В данной статье предлагается исследование RQ-системы ММРР/M/1 с разнотипными вызываемыми заявками методом асимптотически-диффузионного анализа. Получая предельный диффузионный процесс, аппроксимирующий число заявок на орбите, мы решаем уравнение Фоккера-Планка для его плотности и дискретизируем это распределение. Такая аппроксимация достаточно близка к распределению числа заявок на орбите и показывает точные результаты при численных экспериментах.

2. Математическая модель

Рассмотрим однолинейную RQ-систему, на вход которой поступает ММРР-поток заявок. Входящий поток задается инфинитезимальной матрицей управляющего процесса \mathbf{Q} и диагональной матрицей условных интенсивностей Λ . Если прибор свободен поступающие заявки занимают его для обслуживания на случайное время, распределенное экспоненциально с параметром μ_1 .

Если поступившая заявка застает прибор занятым, она отправляется на орбиту, где осуществляет случайную задержку. Длительность задержки также имеет экспоненциальное распределение с параметром σ . После задержки заявка возвращается в систему и вновь пытается занять прибор. Дальнейшее её поведение не отличается от вновь прибывших заявок.

Когда прибор свободен он вызывает заявки извне. Для удобства исследования пронумеруем типы вызываемых заявок от 2 до N . Прибор вызывает заявки типа n с интенсивностью α_n . Вызванная заявка мгновенно занимает прибор для обслуживания. Время обслуживания вызываемой заявки типа n распределено по экспоненциальному закону с параметром μ_n .

Пусть $i(t)$ – число заявок на орбите в момент времени t , а $k(t)$ – состояние прибора в момент времени t . Состояния процесса $k(t)$ могут принимать следующие

значения:

$$k(t) = \begin{cases} 0, & \text{если прибор свободен,} \\ 1, & \text{если обслуживается поступившая заявка,} \\ n, & \text{если обслуживается вызванная заявка типа } n, n = \overline{0, N}. \end{cases}$$

Также введем случайный процесс $m(t)$ – цепь Маркова, управляющую ММРР-потоком, с конечным числом состояний M и непрерывным временем. Когда $m(t) = m$ интенсивность входящего потока равна λ_m , $m = \overline{1, M}$.

Трехмерный процесс $\{i(t), k(t), m(t)\}$ образует цепь Маркова с непрерывным временем. Будем полагать, что цепь Маркова эргодическая и что стационарное распределение вероятностей состояний данного процесса существует. Обозначим $P\{i(t) = i, k(t) = k, m(t) = m\} = P_k(i, m, t)$ – вероятность того, что в момент времени t прибор находится в состоянии k , в системе находится i заявок и управляющая ММРР-потоком цепь Маркова $m(t)$ принимает значение m .

Запишем систему дифференциальных уравнений Колмогорова для частичных характеристических функций $H_k(u, m, t) = \sum_{i=0}^{\infty} e^{ju} P_k(i, m, t)$, $k = \overline{0, N}$, где $j = \sqrt{-1}$, в матричном виде:

$$\begin{aligned} \frac{\partial \mathbf{H}_0(u, t)}{\partial t} &= \mathbf{H}_0(u, t) \left(\mathbf{Q} - \boldsymbol{\Lambda} - \sum_{n=2}^N \alpha_n \mathbf{I} \right) + j\sigma \frac{\partial \mathbf{H}_0(u, t)}{\partial t} + \sum_{k=1}^N \mu_k \mathbf{H}_k(u, t), \\ \frac{\partial \mathbf{H}_1(u, t)}{\partial t} &= \mathbf{H}_1(u, t) (\mathbf{Q} + (e^{ju} - 1) \boldsymbol{\Lambda} - \mu_1 \mathbf{I}) + \mathbf{H}_0(u, t) \boldsymbol{\Lambda} - j\sigma e^{-ju} \frac{\partial \mathbf{H}_0(u, t)}{\partial u}, \\ \frac{\partial \mathbf{H}_n(u, t)}{\partial t} &= \mathbf{H}_n(u, t) (\mathbf{Q} + (e^{ju} - 1) \boldsymbol{\Lambda} - \mu_n \mathbf{I}) + \alpha_n \mathbf{H}_0(u, t), \quad n = \overline{2, N}. \end{aligned} \quad (1)$$

Векторы $\mathbf{H}_n(u, t)$ состоят из частичных характеристических функций расположенных по возрастанию аргумента t , \mathbf{I} – единичная матрица.

Далее построим диффузионную аппроксимацию процесса $i(t)$ на основе полученных уравнений.

3. Первый этап асимптотически-диффузионного анализа

В системе уравнений (1) введем следующие замены:

$$\sigma = \varepsilon, \quad \tau = \varepsilon t, \quad u = \varepsilon w, \quad \mathbf{H}_k(u, t) = \mathbf{F}_k(w, \tau, \varepsilon), \quad k = \overline{0, N},$$

где τ имеет смысл нормированного времени, w – аргумент асимптотической характеристической функции, получим

$$\varepsilon \frac{\partial \mathbf{F}_0(w, \tau, \varepsilon)}{\partial \tau} = \mathbf{F}_0(w, \tau, \varepsilon) \left(\mathbf{Q} - \boldsymbol{\Lambda} - \sum_{n=2}^N \alpha_n \mathbf{I} \right) +$$

$$\begin{aligned}
 & +j \frac{\partial \mathbf{F}_0(w, \tau, \varepsilon)}{\partial w} + \sum_{k=1}^N \mu_k \mathbf{F}_k(w, \tau, \varepsilon), \\
 \varepsilon \frac{\partial \mathbf{F}_1(w, \tau, \varepsilon)}{\partial \tau} & = \mathbf{F}_1(w, \tau, \varepsilon) (\mathbf{Q} + (e^{jw\varepsilon} - 1)\mathbf{\Lambda} - \mu_1 \mathbf{I}) + \\
 & + \mathbf{F}_0(w, \tau, \varepsilon) \mathbf{\Lambda} - j e^{-jw\varepsilon} \frac{\partial \mathbf{F}_0(w, \tau, \varepsilon)}{\partial w}, \\
 \varepsilon \frac{\partial \mathbf{F}_n(w, \tau, \varepsilon)}{\partial \tau} & = \mathbf{F}_n(w, \tau, \varepsilon) (\mathbf{Q} + (e^{jw\varepsilon} - 1)\mathbf{\Lambda} - \mu_n \mathbf{I}) + \\
 & + \alpha_n \mathbf{F}_0(w, \tau, \varepsilon), \quad n = \overline{2, N}.
 \end{aligned} \tag{2}$$

Теорема 1. В предельном условии $\sigma \rightarrow 0$ верно следующее равенство

$$\lim_{\sigma \rightarrow 0} \mathbb{M} e^{jw\sigma i(\frac{\tau}{\sigma})} = e^{jwx(\tau)}, \tag{3}$$

где функция $x(\tau)$ является решением дифференциального уравнения

$$x'(\tau) = -x(\tau) \mathbf{r}_0 \mathbf{e} + (\mathbf{r} - \mathbf{r}_0) \mathbf{\Lambda} \mathbf{e}. \tag{4}$$

Здесь \mathbf{r} – стационарное распределение состояний процесса $m(t)$, \mathbf{e} – единичный вектор, \mathbf{r}_n – двумерное стационарное распределение вероятностей состояний прибора и состояний цепи Маркова, управляющей входящим потоком, удовлетворяющее системе уравнений

$$\begin{aligned}
 \mathbf{r}_0 \left(\mathbf{Q} - \mathbf{\Lambda} - \sum_{n=2}^N \alpha_n \mathbf{I} - x \mathbf{I} \right) + \sum_{k=1}^N \mu_k \mathbf{r}_k & = 0, \\
 \mathbf{r}_1 (\mathbf{Q} - \mu_1 \mathbf{I}) + \mathbf{r}_0 (\mathbf{\Lambda} + x \mathbf{I}) & = 0, \\
 \mathbf{r}_n (\mathbf{Q} - \mu_n \mathbf{I}) + \alpha_n \mathbf{r}_0 & = 0, \quad n = \overline{2, N}, \\
 \sum_{n=0}^N \mathbf{r}_n & = \mathbf{r}.
 \end{aligned} \tag{5}$$

Обозначим

$$a(x) = -x \mathbf{r}_0 \mathbf{e} + (\mathbf{r} - \mathbf{r}_0) \mathbf{\Lambda} \mathbf{e}. \tag{6}$$

Функция $a(x)$ имеет смысл коэффициента переноса диффузационного процесса, аппроксимирующего число заявок на орбите исследуемой системы.

4. Второй этап асимптотически-диффузионного анализа

Центрируем случайный процесс $i(t)$ введя в системе (1) следующие замены

$$\mathbf{H}_k(u, t) = e^{j\frac{u}{\sigma}x(\sigma t)} \mathbf{H}_k^{(2)}(u, t),$$

тогда для частичных характеристических функций центрированного числа заявок на орбите получим систему

$$\begin{aligned} \frac{\partial \mathbf{H}_0^{(2)}(u, t)}{\partial t} + jux'(\sigma t) \mathbf{H}_0^{(2)}(u, t) &= \mathbf{H}_0^{(2)}(u, t) \left(\mathbf{Q} - \mathbf{\Lambda} - \left(x(\sigma t) + \sum_{n=2}^N \alpha_n \right) \mathbf{I} \right) + \\ &\quad + j\sigma \frac{\partial \mathbf{H}_0^{(2)}(u, t)}{\partial u} + \sum_{k=1}^N \mu_k \mathbf{H}_k^{(2)}(u, t), \\ \frac{\partial \mathbf{H}_1^{(2)}(u, t)}{\partial t} + jux'(\sigma t) \mathbf{H}_1^{(2)}(u, t) &= \mathbf{H}_1^{(2)}(u, t) (\mathbf{Q} + (e^{ju} - 1) \mathbf{\Lambda} - \mu_1 \mathbf{I}) + \\ &\quad + \mathbf{H}_0^{(2)}(u, t) (\mathbf{\Lambda} + e^{-ju} x(\sigma t)) - j\sigma e^{-ju} \frac{\partial \mathbf{H}_0^{(2)}(u, t)}{\partial u}, \\ \frac{\partial \mathbf{H}_n^{(2)}(u, t)}{\partial t} + jux'(\sigma t) \mathbf{H}_n^{(2)}(u, t) &= \mathbf{H}_n^{(2)}(u, t) (\mathbf{Q} + (e^{ju} - 1) \mathbf{\Lambda} - \mu_n \mathbf{I}) + \\ &\quad + \alpha_n \mathbf{H}_0^{(2)}(u, t), \quad n = \overline{2, N}, \end{aligned} \tag{7}$$

В полученной системе уравнений вводим замену переменных следующего вида

$$\sigma = \varepsilon^2, \quad \tau = \varepsilon^2 t, \quad u = \varepsilon w, \quad \mathbf{H}_k^{(2)}(u, t) = \mathbf{F}_k^{(2)}(w, \tau, \varepsilon), \quad k = \overline{0, N},$$

что дает нам

$$\begin{aligned} &\varepsilon^2 \frac{\partial \mathbf{F}_0^{(2)}(w, \tau, \varepsilon)}{\partial \tau} + j\varepsilon wa(x) \mathbf{F}_0^{(2)}(w, \tau, \varepsilon) = \\ &= \mathbf{F}_0^{(2)}(w, \tau, \varepsilon) \left(\mathbf{Q} - \mathbf{\Lambda} - \left(x(\tau) + \sum_{n=2}^N \alpha_n \right) \mathbf{I} \right) + j\varepsilon \frac{\partial \mathbf{F}_0^{(2)}(w, \tau, \varepsilon)}{\partial w} + \sum_{k=1}^N \mu_k \mathbf{F}_k^{(2)}(w, \tau, \varepsilon), \\ &\varepsilon^2 \frac{\partial \mathbf{F}_1^{(2)}(w, \tau, \varepsilon)}{\partial \tau} + j\varepsilon wa(x) \mathbf{F}_1^{(2)}(w, \tau, \varepsilon) = \mathbf{F}_1^{(2)}(w, \tau, \varepsilon) (\mathbf{Q} + (e^{ju} - 1) \mathbf{\Lambda} - \mu_1 \mathbf{I}) + \\ &\quad + \mathbf{F}_0^{(2)}(w, \tau, \varepsilon) (\mathbf{\Lambda} + e^{-jw\varepsilon} x(\tau) \mathbf{I}) - j\varepsilon e^{-jw\varepsilon} \frac{\partial \mathbf{F}_0^{(2)}(w, \tau, \varepsilon)}{\partial w}, \\ &\varepsilon^2 \frac{\partial \mathbf{F}_n^{(2)}(w, \tau, \varepsilon)}{\partial \tau} + j\varepsilon wa(x) \mathbf{F}_n^{(2)}(w, \tau, \varepsilon) = \\ &= \mathbf{F}_n^{(2)}(w, \tau, \varepsilon) (\mathbf{Q} + (e^{jw\varepsilon} - 1) \mathbf{\Lambda} - \mu_n \mathbf{I}) + \alpha_n \mathbf{F}_0^{(2)}(w, \tau, \varepsilon). \end{aligned} \tag{8}$$

Теорема 2. *Функции $\mathbf{F}_n^{(2)}(w, \tau, \varepsilon)$ в пределе при $\varepsilon \rightarrow 0$ имеют вид*

$$\lim_{\varepsilon \rightarrow 0} \mathbf{F}_k^{(2)}(w, \tau, \varepsilon) = \mathbf{F}_k^{(2)}(w, \tau) = \Phi(w, \tau) \mathbf{r}_k,$$

где векторы \mathbf{r}_k получены в теореме 1, а $\Phi(w, \tau)$ является характеристической функцией аппроксимирующего случайногопроцесса и удовлетворяет дифференциальному уравнению

$$\frac{\partial \Phi(w, \tau)}{\partial \tau} = w \frac{\partial \Phi(w, \tau)}{\partial w} a'(x) + \frac{(jw)^2}{2} \Phi(w, \tau) b(x). \quad (9)$$

Здесь $b(x) = a(x) + 2 \{-x \mathbf{g}_0 \mathbf{e} + (\mathbf{g} - \mathbf{g}_0) \Lambda \mathbf{e} + x \mathbf{r}_0 \mathbf{e}\}$, векторы \mathbf{g}_n являются решением системы уравнений

$$\begin{aligned} \mathbf{g}_0 \left(\mathbf{Q} - \Lambda - \left(x + \sum_{n=2}^N \alpha_n \right) \mathbf{I} \right) + \sum_{k=1}^N \mu_k \mathbf{g}_k &= a(x) \mathbf{r}_0, \\ \mathbf{g}_0(\Lambda + x \mathbf{I}) + \mathbf{g}_1(\mathbf{Q} - \mu_1 \mathbf{I}) &= a(x) \mathbf{r}_1 - \mathbf{r}_1 \Lambda + x \mathbf{r}_0, \\ \alpha_n \mathbf{g}_0 + \mathbf{g}_n(\mathbf{Q} - \mu_n \mathbf{I}) &= a(x) \mathbf{r}_n - \mathbf{r}_n \Lambda, \quad n = \overline{2, N}, \\ \sum_{k=0}^N \mathbf{g}_k \mathbf{e} &= 0. \end{aligned} \quad (10)$$

5. Аппроксимация распределения вероятностей числа заявок на орбите

Уравнение (9) является преобразованием Фурье от уравнения Фоккера-Планка для плотности $D(y, \tau)$ диффузионного процесса $y(\tau)$, аппроксимирующего центрированное и нормированное число заявок на орбите. Применяя обратное преобразование Фурье получим

$$\frac{\partial D(y, \tau)}{\partial \tau} = -\frac{\partial}{\partial y} \{a'(x)yD(y, \tau)\} + \frac{1}{2} \frac{\partial^2}{\partial y^2} \{b(x)D(y, \tau)\}. \quad (11)$$

Введем случайный процесс $z(\tau)$ вида

$$z(\tau) = x(\tau) + \varepsilon y(\tau), \quad (12)$$

для плотности которого $S(z)$ запишем уравнение Фоккера-Планка с учетом (4) и (11)

$$S(z) = \frac{C}{b(z)} \exp \left\{ \frac{2}{\sigma} \int_0^z \frac{a(x)}{b(x)} dx \right\}, \quad (13)$$

где C – нормирующий множитель. Тогда можем записать формулу для аппроксимации $P_{Diff}(i)$ распределения вероятностей числа заявок на орбите

$$P_{Diff}(i) = \frac{S(\sigma i)}{\sum_{n=0}^{\infty} S(\sigma n)}. \quad (14)$$

6. Заключение

В работе представлен процесс построения диффузионной аппроксимации числа заявок на орбите RQ-системы ММРР/М/1 с разнотипными вызываемыми заявками. Приведены формулы коэффициентов переноса и диффузии аппроксимирующего процесса, а также формула для аппроксимации дискретного распределения вероятностей числа заявок на орбите.

Литература

1. S. Aguir, F. Karaesmen, O. Z. Aksin, F. Chauvet, The impact of retrials on call center performance, *OR Spectrum* 26 (3) (2004) 353–376.
2. A. Kuki, J. Sztrik, A. Toth, T. Berczes, A contribution to modeling two-way communication with retrial queueing systems, in: *Information Technologies and Mathematical Modelling. Queueing Theory and Applications*, Springer, 2018, pp. 236–247.
3. J. Sztrik, A. Toth, A. Pinter, Z. Bacs, Simulation of finite-source retrial queues with two-way communications to the orbit, in: *International Conference on Information Technologies and Mathematical Modelling*, Springer, 2019, pp. 270–284.
4. A. Toth, J. Sztrik, A. Kuki, T. Berczes, D. Effosinin, Reliability analysis of finite-source retrial queues with outgoing calls using simulation, in: *2019 International Conference on Information and Digital Technologies (IDT)*, IEEE, 2019, pp. 504–511.
5. V. Dragieva, T. Phung-Duc, Two-way communication m/m/1 retrial queue with server-orbit interaction, in: *Proceedings of the 11th International Conference on Queueing Theory and Network Applications*, 2016, pp. 1–7.
6. E. Morozov, T. Phung-Duc, Regenerative analysis of two-way communication orbit-queue with general service time, in: *International Conference on Queueing Theory and Network Applications*, Springer, 2018, pp. 22–32.
7. S. Ouazine, K. Abbas, A functional approximation for retrial queues with two way communication, *Annals of Operations Research* 247 (1) (2016) 211–227.

UDC: 51-74, 004.048, 004.93

Network traffic preparation for its states analysis by the aggregated data packets partial correlations method

Nikol'skii D.N.¹ and Krasnov A.E.²

¹Smart Wallet LLC (SWiP), Nezhinskaya str., 1, bldg. 4, 159, Moscow, Russia

²Russian State Social University, Wilhelm Peak str., 4, bldg. 1, Moscow, Russia

nikolskydn@mail.ru, krasnovmgutu@yandex.ru

Abstract

The aggregation procedures network traffic data packets for the analysis of its states by the partial correlations method of aggregated data packets are described. The headers of network traffic data packets parameters transformation to time series of its aggregated data packets are presented. The novelty of the approach lies in the fact that the streams of network traffic aggregated data packets are formed considering bit flags from the TCP header of the TCP/IP data transmission protocol. The approach has shown high efficiency for identifying various types of DDoS attacks.

Keywords: network traffic, TCP/IP protocol, TCP header bit flag, data packet aggregation, partial correlations, DDoS attacks identification.

1. Introduction

DDoS attacks are complex types of attacks [1] and special methods of network traffic analysis are required to identify (detect and classify) them. Such methods based on network traffic data packets aggregation and subsequent partial correlations calculation of adjacent aggregates were developed and explored by the authors in [2, 3]. The high efficiency identification of various types of DDoS attacks based on the non-stationary states description of network traffic by the so-called patterns — histograms of the values of partial correlations of adjacent aggregates is shown [3]. At the same time, in these works, no attention was paid to the description of the procedures for the formation of the aggregates themselves. Papers [4, 5, 6] are devoted to the formation of network traffic aggregates for various purposes.

The purpose of this work is a detailed description of the procedures for aggregating network traffic data packets for analyzing its states by the partial correlations method of aggregated data packets. The specific tasks of the work are: collection of network

traffic — description of the transformation of the parameters of the headers of its data packets; preparation of network traffic — formation time series of its aggregated packets, considering bit flags from the data transmission protocol TCP header.

2. Network traffic collection

The network traffic primary signs were represented by the TCP/IP data transfer protocol [7, 8]. Bit flags from the TCP network packet header were selected for the study. The change in the statistical characteristics of these flags over time can be used to describe the unique network traffic states: its normal and abnormal states.

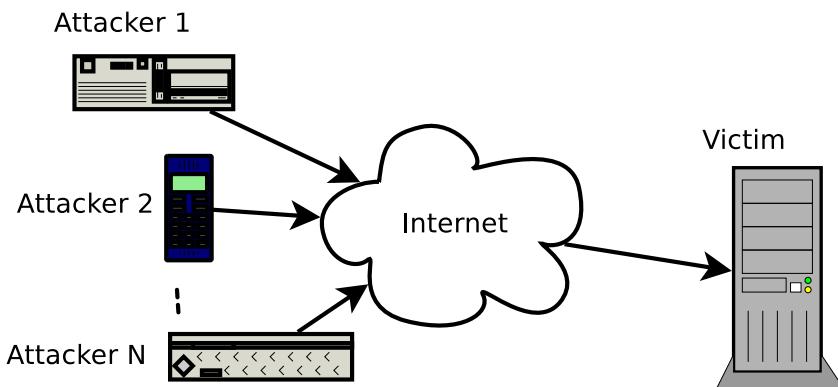


Fig. 1. Experimental stand for picking up network traffic

To study the traffic, a stand was created (Fig. 1), using which attacks were carried out on the victim node from distributed nodes. The some Frontend server of some Web-service was chosen as a victim node. A Qlogic 10 Gb/s network card was used as hardware support. The packages were parsed with the tshark utility. As a result, a csv file was obtained with network packets stored on the victim node.

As an example, in table 1 first 10 packets on the victim host when performing a Slowloris attack are shown. The columns are:

- `frame.time` is the packet saving time, the start of the report is taken as the time when the first packet fell;
- `ip.len` is the total length of the packet in bytes;
- `tcp.flags` TCP flags.

3. Network traffic cooking

To build time series (a aggregated data stream packages with different flag indices), table 1 was pivoted. The new table columns are the flags values from the

Number	frame.time, sec.	tcp.flags	ip.len, byte
0	0.00000	0x10	52
1	0.00004	0x04	40
2	0.00494	0x02	60
3	0.00497	0x12	60
4	0.00528	0x10	52
5	0.01012	0x02	60
6	0.01014	0x12	60
7	0.01021	0x18	286
8	0.01023	0x10	52
9	0.01047	0x10	52

Table 1. Captured network packets

TCP headers `tcp.flags`. The values in these columns are the values sum from the `ip.len` field. The data pivoting result with source packets for a Slowloris attack are presented in table 2

frame.time \ tcp.flags	0x02	0x04	0x10	0x11	0x12	0x18
0 days 00:00:00	0.0	0.0	52.0	0.0	0.0	0.0
0 days 00:00:00.000040	0.0	40.0	0.0	0.0	0.0	0.0
0 days 00:00:00.004940	60.0	0.0	0.0	0.0	0.0	0.0
0 days 00:00:00.004970	0.0	0.0	0.0	0.0	60.0	0.0
0 days 00:00:00.005280	0.0	0.0	52.0	0.0	0.0	0.0
0 days 00:00:00.010120	60.0	0.0	0.0	0.0	0.0	0.0
0 days 00:00:00.010140	0.0	0.0	0.0	0.0	60.0	0.0
0 days 00:00:00.010210	0.0	0.0	0.0	0.0	0.0	286.0
0 days 00:00:00.010230	0.0	0.0	52.0	0.0	0.0	0.0
0 days 00:00:00.010470	0.0	0.0	52.0	0.0	0.0	0.0

Table 2. Packages after pivoting

Next, the received data aggregation and the formation of a time stream were performed: changes in the aggregated data packets count and their total length over time. As a result, for each j -th flag index ($j = 1, 2, \dots, J$) of each aggregate of the interval ΔT , the number of packets $N_{\Delta T}^{(j)}(t)$ and their total length or information capacity $I_{\Delta T}^{(j)}(t)$ ($j = 1, 2, \dots, J$) at the current time t .

The pictures below show the first 30 seconds of the process. So, in Fig. 2, aggregation is performed using the function of counting the number of packets, and in Fig. 3 using the function of counting the total length of packets. The size ΔT of the aggregation window is equal to one second.

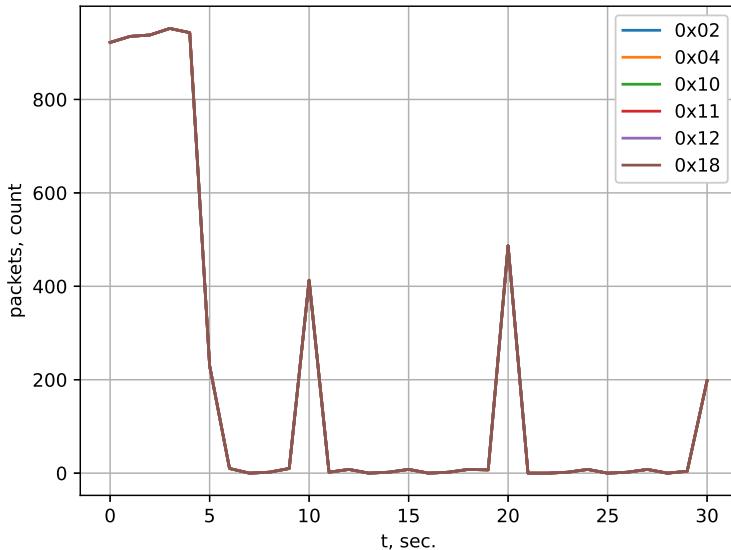


Fig. 2. Changes in the number of packets over time

4. Conclusion

The procedures described above make it possible to form a network traffic aggregates discrete flow model in the form of analytical signals [3]:

$$F_{\Delta T}^{(j)}(t_k) = X_{\Delta T}^{(j)}(t_k) + iY_{\Delta T}^{(j)}(t_k), \quad j = 1, 2, \dots, J, \quad k = 0, 1, \dots, K,$$

characterized by dynamic coordinates $X_{\Delta T}^{(j)}(t_k)$ and velocities $Y_{\Delta T}^{(j)}(t_k)$, which are related to the number $N_{\Delta T}^{(j)}(t_k)$ of packets in the k -th aggregates, the progress of their birth/destruction, and flag indices j . The values $I_{\Delta T}^{(j)}(t_k)$ of the lengths of the aggregates make it possible to introduce their statistical weights $I_{\Delta T}^{(j)}(t_k) / \sum_{j=1}^J X_{\Delta T}^{(j)}(t_k)$ or the significance of the flag indices j .

REFERENCES

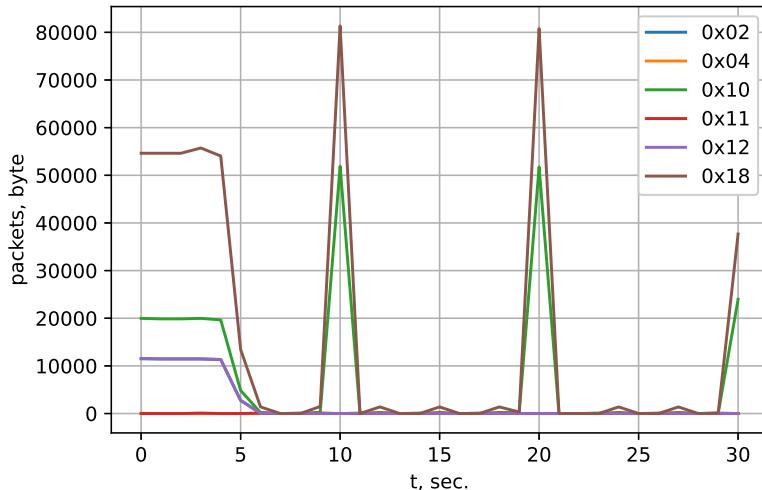


Fig. 3. Changes in the total length of packets over time

1. Bhattacharyya D., Kalita J. DDoS attacks: evolution, detection, prevention, reaction, and tolerance. Boca Raton: CRC Press Taylor & Francis Group, 2016.
2. Krasnov A.E., Nikol'skii D.N. Statistical Distributions of Partial Correlators of Network Traffic Aggregated Packets for Distinguishing DDoS Attacks // Distributed Computer and Communication Networks. DCCN 2019. Lecture Notes in Computer Science. Springer, Cham. 2019. V. 11965. P. 365–378.
3. Krasnov A.E., Nikol'skii D.N. Formation of One-Dimensional Distributions of Values of Correlators of Network Traffic Aggregates // Russian Physics Journal. 2020. V. 63. P. 563–573.
4. Konoplev V.V., Zaharov D.YU., Boyarskij M.N., Nazirov R.R. Skhema adaptivnogo agregirovaniya dlya klasterizacii dannyh setevogo trafika (in Russia) // "ISSLEDOVANO V ROSSI". 2003. P. 2555–2568.
5. Get'man A.I., Ivannikov V.P., Markin YU.V., Padaryan V.A., Tihonov A.YU. Model' predstavleniya dannyh pri provedenii glubokogo analiza setevogo trafika (in Russia) // Trudy ISP RAN. 2015. T.27, N. 4. S. 5–20.
6. Efimov A.YU. Problemy obrabotki statistiki setevogo trafika dlya obonaruzheniya vtorzhenij v sushchestvuyushchih informacionnyh sistemah (in Russia) // Programmnye produkty i sistemy / Software & Systems. 2016. N. 1. S. 17–21.

7. Postel J. (ed.). Internet protocol - DARPA Internet Program Protocol Specification, rfc791 (Internet Standard), Los Angeles: USC/Information Sciences Institute, 1981.
8. Postel J. (ed.). Transmission control protocol - DARPA Internet Program Protocol Specification, rfc793 (Internet Standard), Los Angeles: USC/Information Sciences Institute, 1981.

UDC: 123.456

Stability analysis of an unreliable two-class retrial system with constant retrial rates

R. S. Nekrasova^{1,2}, E. V. Morozov^{1,2,3}, D. V. Efrosinin^{4,5}

¹IAMR Karelian Research Centre RAS, Pushkinskaay 11, Petrozavodsk, Russia

²Petrozavodsk State University, Lenin 33, Petrozavodsk, Russia

³Moscow Center for Fundamental and Applied Mathematics, Moscow State University, Moscow 119991, Russia

⁴Johannes Kepler University, Linz, Austria

⁵Peoples Friendship University of Russia, Moscow Russia

ruslana.nekrasova@mail.ru, emorozov@karelia.ru, dmitry.efrosinin@mail.ru

Abstract

A two-class single-server retrial system with Poisson inputs is considered. In this system, unlike conventional retrial systems, each new class- i customer joins the “end” of a virtual class- i orbit, and the “oldest” customer from each orbit is only allowed to make an attempt to occupy server after a class-dependent exponential retrial time. Moreover, the server is assumed to be not reliable, and a customer whose service is interrupted joins the “top” of class- i orbit queue. Thus FIFO discipline is applied in both orbits. Using regenerative methodology we derive stability conditions of this system.

Keywords: retrial system, unreliable server, regenerative stability analysis

1. Introduction

It is well known that the models of the queueing theory are widely used to analyze the operational properties of telecommunication and computer service systems, and also help in solving various optimization problems. One of the most important element of such analysis is the study of the steady-state behaviour of the system, when the performance and reliability metrics as well as the optimal control policies in controllable models are independent of time. However, in order to obtain results it is necessary to guarantee the existence of a stationary mode, i.e. to obtain some ergodicity or stability condition for the corresponding queueing systems. A large variety of queues with derived ergodic conditions can be found in numerous papers and monographs, e.g. in [1, 3, 4, 8] and references cited therein.

*

Although many papers have been devoted to the stability of queues, there are still gaps for certain classes of systems which are characterized by different additional features like reliability attributes of a server, retrial phenomenon and different classes of customers. The server is unreliable when it is subject to breakdowns and repairs. The retrial effect occurs in systems where a primary customer is blocked when the server is busy. In this case the customer joins the queue of retrial customers where it repeats the attempt to occupy the server either independently of other orbiting customers due to the classical retrial discipline or only if it is located at the head of the retrial queue with respect to the constant retrial discipline. Different classes of customers arise in situations where a certain group of customers is willing to incur additional costs to reduce waiting times and improve service quality. The existence of two classes of customers is the motivation for considering the optimal scheduling problem, which cannot be solved without an accurate understanding of the stability conditions of a given system. For example, there are still relatively few works on ergodicity condition of single-server systems combined with double retrial queues for different classes of customers. The stability of the system consisting of ordinary and executive orbit with classical retrial discipline was investigated in [5]. In [6], the authors proposed ergodicity condition for the retrial queueing system with two orbits operating under a constant retrial discipline under specified symmetric constraints when arrival, retrial and service rates are equal for different classes of customers. The stability analysis under the same retrial discipline but with asymmetric assumptions on the rates was provided in [2]. Stability conditions for some other double orbit retrial queues with classic retrial discipline with different modifications were established in [9, 10], where the service process was accompanied with interruptions.

The model discussed in this article differs from those previously studied. The queueing system combines unreliable server, involves two classes of customers and the repeated customers with constant retrial rate. All the customers without exception are served in order of arrival time, i.e. according to the FIFO rule. In other words, the primary arrival has not an access to the server independently of its state and it becomes upon arrival immediately blocked. This customer must be sent straight to the end of the retrial queue where it waits until it is at the head of the queue where it attempts to occupy the server. Moreover, the system is assumed to be asymmetric, i.e. different classes of customers prescribe unequal arrival, retrial and service intensities.

2. Description of the model

We consider a two-class single-server retrial system with constant retrial rate in which the arrivals of class i follow a Poisson input with rate λ_i , $i = 1, 2$. We denote the system by Σ , and let $\lambda = \lambda_1 + \lambda_2$ denote the total input rate.

The model Σ has the following important property. Each new class- i customer joins the 'end' of a virtual class- i orbit (even if the server is idle upon his arrival). The customers in orbit i form a FIFO-type virtual queue. In other words, only the top ('oldest') customer in orbit i makes attempts to occupy server after exponential (retrial) time ξ_i . We denote the rate of class- i orbit by $\theta_i = 1/\mathbb{E}\xi_i$. Thus the retrial rate of the orbital customers stays constant independently on orbit size (the number of customers in orbit), and the model belongs to a special class of the retrial queuing system with a *constant retrial rate*. The orbital customer makes the attempts until he finds the server idle and ready for service. In general the server is unreliable. Namely, during service of a class- i customer a failure of server happens with rate $\alpha_i := 1/\mathbb{E}A_i$ where A_i is class- i failure time. After a failure the server becomes ready to work after exponential repair time B with rate $\beta = 1/\mathbb{E}B$. It is worth mentioning that the repair time is assumed to be independent of the class number. If the service of a customer is interrupted by a failure then the customer returns to the 'top' of the corresponding virtual orbit-queue. Thus we use classic FIFO discipline. The service times are assumed to be exponential and class-dependent with the corresponding service rate μ_i . Finally, we define the class- i traffic intensity

$$\rho_i = \lambda_i / \mu_i, \quad i = 1, 2.$$

3. Stability analysis

Our approach to stability analysis of this system is based on the technique developed in [7] to find stability criterion of two-dimensional Markov Chains. We outline the approach. Let $Y^{(i)}(t)$ be the number of the customers in orbit i at instant t . Consider the sequence $\{D_n, n \geq 1\}$ of instants, when the server become idle and ready for the service either after the competition of previous service or because the repair is finished. Introduce the embedded (discrete-time) orbit size process,

$$Y_n^{(i)} := Y^{(i)}(D_n^+), \quad i = 1, 2, n \geq 1,$$

which describes the state of orbit i just after the departure instant D_n . Because the governing distributions are assumed to be exponential, then it is easy to see that the process

$$\mathbf{Y} = \{Y_n^{(1)}, Y_n^{(2)}\}, \quad n \geq 1,$$

defines a two-dimensional Markov Chain. It is evident that the ergodicity of the Markov Chain \mathbf{Y} means the stability of the system under consideration. In the book

[7] the ergodicity criterion of a two-dimensional Markov Chain in an explicit form have been obtained. More precisely, this criterion is the set of the following *negative drift conditions* expressed via the mean increments of the components of the Markov Chain:

$$\begin{cases} M_1^{11}M_2^{10} - M_2^{11}M_1^{10} < 0, \\ M_2^{11}M_1^{01} - M_1^{11}M_2^{01} < 0, \end{cases} \quad (1)$$

where

$$M_i^{11} = E[Y_{n+1}^{(i)} - Y_n^{(i)} | Y_n^{(1)} > 0, Y_n^{(2)} > 0], \quad (2)$$

$$M_i^{01} = E[Y_{n+1}^{(i)} - Y_n^{(i)} | Y_n^{(1)} = 0, Y_n^{(2)} > 0], \quad (3)$$

$$M_i^{10} = E[Y_{n+1}^{(i)} - Y_n^{(i)} | Y_n^{(1)} > 0, Y_n^{(2)} = 0], \quad i = 1, 2, \quad (4)$$

represent the conditional drifts of the components of this Markov Chain between the departures of the customers leaving the system.

We note that in a recent paper [2] this approach is applied to establish stability criterion of a conventional two-class retrial system with constant retrial rate and reliable server, where the incoming customers follow Poisson inputs and immediately receive the service if meet server idle.

Now return to the model we study in the present paper and denote $\rho_1 = \lambda_1/\mu_1$, $\rho_2 = \lambda_2/\mu_2$. If we replace the inequalities (1) by the equations using an explicit form of the (conditional) drifts, then we obtain the following *non-zero* solution (θ_1, θ_2) :

$$\begin{aligned} \theta_1^* &= \frac{\beta\rho_1(\alpha_1 + \mu_1)}{(1 - (\rho_1 + \rho_2))\beta - (\rho_1\alpha_1 + \rho_2\alpha_2)}, \\ \theta_2^* &= \frac{\beta\rho_2(\alpha_2 + \mu_2)}{(1 - (\rho_1 + \rho_2))\beta - (\rho_1\alpha_1 + \rho_2\alpha_2)}, \end{aligned}$$

which is *positive* if condition

$$(\rho_1 + \rho_2) + \frac{\rho_1\alpha_1 + \rho_2\alpha_2}{\beta} < 1 \quad (5)$$

holds. Thus, under condition (5), $\theta_1^* > 0$, $\theta_2^* > 0$, and then one can show that stability zone of the system is *non-empty*. Hence to guarantee that stability region is not empty the values of parameters λ_i , μ_i , α_i and β must satisfy condition (5).

To illustrate the obtained results, consider a particular case of the symmetric classes

$$\lambda_1 = 1, \lambda_2 = 1, \mu_1 = 3, \mu_2 = 3, \alpha_1 = 2, \alpha_2 = 2, \beta = 5.$$

Thus $\theta_1^* = \theta_2^* = 25$. Then the system (both orbits) is stable (ergodic) if and only if inequalities (1) are satisfied which in this case take the following form:

$$\begin{cases} \theta_2(-3\theta_1^2 + 107\theta_1 + 75) + \theta_1(-3\theta_2^2 + 32\theta_2 + 220) < 0, \\ \theta_1(-3\theta_2^2 + 107\theta_2 + 75) + \theta_2(-3\theta_1^2 + 32\theta_1 + 220) < 0. \end{cases} \quad (6)$$

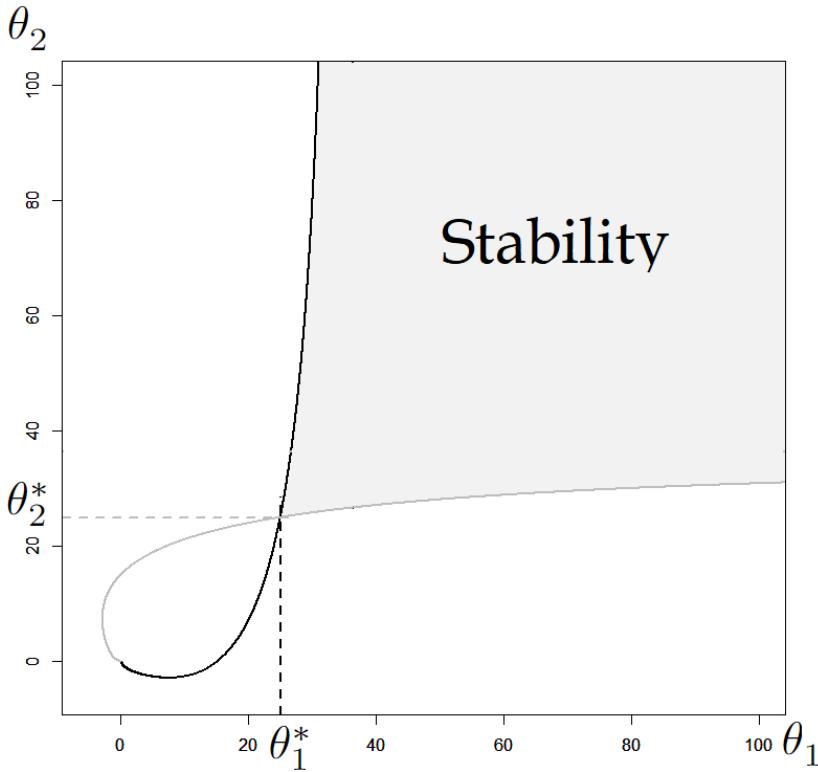


Fig. 1. Stability zone for retrial rates in symmetrical class model.

The stability region for this example is presented by a 'grey zone' on Figure 1.

4. Conclusion

Using the regenerative methodology, we study the stability of a two-class retrial system with constant retrial rates and with unreliable server. Based on a fundamental results from the book [7] we derive the stability criterion of the two-dimensional Markov chain describing the dynamics of this system. This result is illustrated by

a numerical example related to the system with the symmetric classes. In a future research we aim to verify the accuracy of condition (5) by simulation and obtain stability conditions for some particular configurations of the system in an explicit form. Moreover it is assumed to extend the analysis to the system with service times with general distributions.

REFERENCES

1. Artalejo, J. R., Gomez-Corral, A.: Retrial queueing systems. Springer. (2008)
2. Avrachenkov, K, Morozov, E., Nekrasova, R.: Stability analysis of two-class retrial systems with constant retrial rates and general service times. ArXiv, abs/2110.09840 (2021)
3. Bramson, M.: Stability of queueing networks. *Probab. Surv.* 5:165–345 (2008)
4. Cohen, J. W.: Analysis of random walks. Amsterdam: I.O.S. Press. (1992)
5. Dimitriou, I.: A queueing model with two classes of retrial customers and paired services. *Ann. Oper. Res.* 238:123–143 (2016)
6. Dimitriou, I.: A two-class queueing system with constant retrial policy and general class dependent service times. *Eur. J. Oper. Res.* 270: 1063–1073 (2018)
7. Fayolle, G., Malyshev, V., Menshikov, M.: Topics in the Constructive Theory of Countable Markov Chains. 1st edn. Cambridge University Press (1995)
8. Gross, D., Shortle, J.F., Thompson J.M., Harris, C.M.: Fundamentals of Queueing Theory, 4th edn. John Wiley & Sons Inc, New Jersey. (2008)
9. Jain, M., Sanga, S.S.: Unreliable single server double orbit retrial queue with balking. *Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci.* 91, 257–268 (2021)
10. Sanga, S.S., Jain, M.: (2019) FM/FM/1 double orbit retrial queue with customers' joining strategy: a parametric nonlinear programming approach. *Appl Math Comput* 362:124542 (2019)

УДК: 519.218.4

О методе моделирования случайной величины с помощью её интенсивности

Г.А. Зверкина¹ and А.А. Кошлев^{1,2}

¹Институт проблем управления им. В. А. Трапезникова Российской академии наук, ул. Профсоюзная, 65, Москва, РФ

²Московский государственный университет имени М.В. Ломоносова, Ленинские горы, д. 1, Москва, РФ

zverkina@gmail.com, koshelev030698@yandex.ru

Аннотация

Одним из самых известных и простых методов моделирования случайных величин является метод обратной функции, основанный на использовании функции, обратной к функции распределения. Однако довольно часто при моделировании поведения нетривиальных моделей исследователь сталкивается с проблемой построения функции, обратной к функции распределения исследуемой случайной величины, а также изменением параметров моделируемых случайных величин. В данном тексте предлагается новый метод случайной величины по одной из определяющих её характеристик – функции интенсивности, даются теоретические обоснования данного метода и его алгоритм.

Ключевые слова: интенсивность случайной величины, численное моделирование, вопросы оптимизации моделирования

1. Введение

Значительное количество известных моделей в теории массового обслуживания (ТМО), в теории надёжности и теории сетей массового обслуживания описывается линейчатыми процессами (см. [4]). При этом стохастические процессы в ТМО и в смежных задачах, как правило, являются эргодическими, т.е. с течением времени распределение такого процесса слабо сходится к некоторому предельному распределению. Однако не всегда можно установить или оценить скорость сходимости распределения эргодического процесса к стационарному распределению, равно как и определить характеристики этого стационарного распределения. Поэтому для получения оценок скорости сходимости распределения

Работа выполнена при частичной финансовой поддержке РФФИ, грант №20-01-00575А. Авторы признательны М.П. Фархадову за ценное обсуждение работы, а также В.В. Козлову за ценные советы и замечания, способствовавшие существенному улучшению содержания текста.

исследуемой системы к стационарному распределению и оценок стационарного распределения может быть использовано *имитационное моделирование* поведения системы с помощью программных средств.

Для этого необходимо уметь моделировать случайные величины с заданным в описании исследуемой модели распределением.

Одним из самых распространенных методов моделирования случайных величин является метод обратной функции или замены переменной [3]: случайная величина ξ с функцией распределения $F(x) \stackrel{\text{def}}{=} \mathbf{P}\{\xi \leq x\}$ может быть промоделирована как

$$\xi \stackrel{\text{def}}{=} F^{-1}(\mathcal{U}), \quad (1)$$

где

$$F^{-1}(y) \stackrel{\text{def}}{=} \inf\{x \in \mathbb{R} : F(x) \geq y\}, \quad y \in \mathbb{R},$$

а \mathcal{U} – равномерно распределённая случайная величина на интервале $[0; 1]$.

При этом имитационное моделирование непрерывных случайных величин фактически является моделированием дискретных случайных величин (аппроксимирующих заданное распределение) – с шагом, определяемым используемым программным продуктом.

Известно, что в некоторых случаях многокомпонентные процессы имеют вложенные процессы восстановления зависящие друг от друга. Например, так может быть в теории надёжности или теории массового обслуживания. В этом случае интенсивность восстановления (отказа/окончания обслуживания/прихода требования) может зависеть от *общего состояния системы* и меняться с течением времени. При этом весь процесс может рассматриваться как марковский и при определенных условиях может являться эргодическим. В этих случаях иногда можно построить грубые оценки скорости сходимости, но для уточнения скорости сходимости полезно применять имитационное моделирование с помощью интенсивности.

Итак, в некоторых сложных стохастических системах “поведение” элементов системы может существенно зависеть от общего состояния системы (т.е. состояний других элементов системы) и не может быть определено до начала моделирования поведения сложной системы. Например, соответствующие случайные величины (такие, как, например, длина периода восстановления) можно определять как случайную величину, заданную с помощью интенсивности окончания того или иного периода, меняющейся в зависимости от полного состояния изучаемой системы (см., например, [11, 12, 10, 9, 13, 14, 2, 15]).

В данной работе представлен алгоритм моделирования случайной величины с помощью её *интенсивности*.

2. Случайные величины и интенсивности

Напомним, как определяется интенсивность в теории восстановления (см. [7]).

Предположим, что продолжительность некоторого периода (работы / обслуживания/ремонта и пр.) определяется поведением *всей* исследуемой системы.

Это можно определить так. Если в момент времени t известно, что еще не закончился период работы (ремонта, ожидания появления нового требования и пр.) и если распределение этого периода ξ имеет *непрерывную* функцию распределения $F(s)$, то вероятность того, что на бесконечно малом промежутке времени $(t, t + \Delta t)$ произойдет окончание интересующего нас периода, имеет вид

$$\begin{aligned} & \mathbf{P}\{\xi \in (t, t + \Delta t) \mid \xi > t\} = \\ & = \frac{\mathbf{P}\{\xi < t + \Delta t\} - \mathbf{P}\{\xi < t\}}{\mathbf{P}\{\xi > t\}} = \frac{F(t + \Delta t) - F(t)}{1 - F(t)} \end{aligned}$$

При $\Delta t \rightarrow 0$ имеем

$$\mathbf{P}\{\xi \in (t, t + \Delta t) \mid \xi > t\} = \lambda(t)\Delta t + o(\Delta t), \quad (2)$$

где

$$\lambda(t) = \frac{p(t)}{1 - F(t)} = \frac{F'(t)}{1 - F(t)} = -\frac{d}{dt} \ln(1 - F(t)),$$

$F(t)$ — функция распределения, $p(t)$ - плотность распределения. Функция $\lambda(t)$ называется интенсивностью случайной величины (см., например, [8]). Мы предполагаем, что интенсивность может зависеть от времени и от некоторых дополнительных параметров, связанных с поведением других элементов сложной системы. Заметим, что

$$F(s) = 1 - \exp\left(-\int_0^s \lambda(u)du\right), \quad p(s) = \lambda(s) \exp\left(-\int_0^s \lambda(u)du\right)$$

Отсюда видно, что п.в. (почти всюду) положительная случайная величина с абсолютно непрерывным распределением, являющаяся периодом некоторого процесса восстановления, определяется однозначно либо функцией распределения, либо плотностью, либо интенсивностью – подробнее см. [7, 1].

3. Предлагаемый метод моделирования случайных величин

Для моделирования поведения сложных систем (массового обслуживания, надёжности и пр.), поведение которых определяется кусочно-линейными процессами, будем использовать не “классический” метод моделирования случайных

величин по схеме (1), а метод, основанный на знании интенсивности окончания соответствующего периода (2).

Рассмотрим последовательность независимых одинаково распределённых величин $\{\mathcal{U}_i\}$, равномерно распределённых на $[0; 1]$.

Время с момента $t = 0$ до произвольного достаточно большого времени T разбьём на интервалы точками $0 < \tau_1 < \tau_2 < \dots < \tau_n = T$. В каждый момент времени $t = \tau_i < T$ предполагается, что известно время пребывания некоторого параметра рассматриваемой сложной системы (обслуживания, надёжности и пр.) в данном состоянии равно величине $s \geq 0$; интенсивность окончания этого периода равна $\lambda(s)$. Тогда в соответствии с (2) изменение этого состояния (т.е. окончание пребывания в данном состоянии) на интервале времени $(t; t + \Delta t)$ (где $\Delta t \stackrel{\text{def}}{=} \tau_{i+1} - \tau_i$) близко к значению $\lambda(s)\Delta t$. Поэтому с вероятностью, равной примерно $\mathbb{P}_s \stackrel{\text{def}}{=} \mathbf{P}\{\lambda(s)\Delta t > \mathcal{U}_i\}$, в момент τ_{i+1} закончится рассматриваемый период времени пребывания исследуемого параметра в данном состоянии.

То есть общая схема моделирования случайной величины по её интенсивности выглядит так:

- 1) Выбираем точки $0 < \tau_1 < \tau_2 < \dots < \tau_n = T$.
- 2) Если выполнено неравенство $\lambda(0)(\tau_1 - 0) > \mathcal{U}_0$, то принимаем, что $\xi = 0$; если это не так, то переходим к следующему этапу.
- 3) Если выполнено неравенство $\lambda(\tau_1)(\tau_2 - \tau_1) > \mathcal{U}_1$, то принимаем, что $\xi = \tau_1$; если это не так, то переходим к следующему этапу.
- 4) Повторяем описанную в п.3 процедуру для τ_i последовательно до тех пор, пока не случится событие

$$\lambda(\tau_i)(\tau_{i+1} - \tau_i) > \mathcal{U}_i.$$

Полагаем $\xi = \tau_i$. Напомним, что $\mathbf{P}\{\mathcal{U} < a\} = a$ при $a \in [0; 1]$.

Легко видеть, что для любой сл.в. ξ с конечным математическим ожиданием эта процедура конечна с вероятностью 1. Надо сказать, что и при

$$\int_0^\infty \lambda(s) ds = +\infty, \quad (3)$$

вышеописанная процедура конечна с вероятностью 1, но её применение нецелесообразно только в случае $\mathbb{E}\xi < \infty$.

Данный метод удобнее классического (формула (1)) тем, что можно моделировать положительные случайные величины независимо от того, насколько сложно (и возможно) вычисление обратной функции к функции распределения случайной

величины. Например, для функции интенсивности $\lambda(s) = \frac{C}{\sqrt[3]{s^3 + 1}}$ невозможно выписать функцию распределения, поскольку интеграл $\int \frac{C}{\sqrt[3]{s^3 + 1}} ds$ не вычисляется в явном виде, но сама подынтегральная функция соответствует формуле (3), т.е. задаёт собственную функцию распределения сл.в.

Авторами были промоделированы различные случайные величины, результат моделирования удовлетворителен и представлен в [6]. Однако, в ходе моделирования было отмечено, что качество моделирования существенно зависит от соотношения абсолютной величины интенсивности $\lambda(\tau_i)$ и шага моделирования $\Delta\tau_i \stackrel{\text{def}}{=} \tau_{i+1} - \tau_i$.

Литература

1. Боровков А. А. *Вероятностные процессы в теории массового обслуживания*. М., 1972.
2. Веретенников А.Ю. *О скорости сходимости к стационарному распределению в системах обслуживания с одним прибором* // Автоматика и телемеханика, 2013, №10, стр. 23-35. Engl. transl. On the rate of convergence to the stationary distribution in the single-server queueing system, Autom. Remote Control 74(10), 1620-1629 (2013).
3. Голенко Д. И. Моделирование и статистический анализ псевдослучайных чисел на электронных вычислительных машинах. М.: Наука, ГРФМЛ, 1969.
4. Гнеденко Б. В., Коваленко И.Н. Введение в теорию массового обслуживания, М., 1966.
5. Зверкина Г. А. Об экспоненциальной сходимости коэффициента готовности // Управление большими системами. 2021. вып. 90. С. 5–35.
6. Зверкина Г. А., Кошелев А. А. Об имитационном моделировании случайных величин с помощью интенсивности // Управление большими системами: сборник трудов. 2021. Вып. 94. С. 33-49.
7. Кокс Д., Смит В. Теория восстановления. М., 1967.
8. Тартаковский А. Г. О последовательном оценивании интенсивности процессов восстановления. // Пробл. передачи информ., 21:1 (1985), 40–47; Problems Inform. Transmission, 21:1 (1985), 30–36
9. Veretennikov A. On convergence rate for Erlang–Sevastyanov type models with infinitely many servers // Theory of Stochastic Processes. 2017. No. 1, 88-102.
10. Veretennikov, A. On mean-field GI/GI/1 queueing model: existence and uniqueness // ANALYTICAL AND COMPUTATIONAL METHODS IN PROBABILITY THEORY AND ITS APPLICATIONS (ACMPT-2017),

- Proceedings of the International Scientific Conference 23–27 October 2017, Under the general editorship of D.Sc. A.V. Lebedev, RUDN, Moscow, Russia, 182-186.
11. Veretennikov, A. On mean-field (GI/GI/1) queueing model: existence and uniqueness // Queueing Syst 2020, 94(3), 243-255.
 12. Veretennikov, A. On Polynomial Recurrence for Reliability System with a Warm Reserve // Markov Processes and Related Fields. 2019. Vol. 25. P. 745-761.
 13. Veretennikov, A. On Recurrence and Availability Factor for Single-Server System With General Arrivals // Reliability: Theory and Applications (RT& A), 2016, vol.11, #3(42), 49-58.
 14. Veretennikov, A. Yu. On the rate of convergence for infinite server Erlang–Sevastyanov’s problem // Queueing Systems, 2014, Volume 76, Issue 2, pp 181-203.
 15. Veretennikov, A. Yu., Zverkina, G. A. Simple Proof of Dynkin’s Formula for Single-Server Systems and Polynomial Convergence Rates // Markov Processes Relat. Fields 20 (2014), 479–504.

UDC: 51.74

On the analysis of a resource loss system with the waiting buffer

A.V. Daraseliya¹ and E.S. Sopin^{1,2}

¹Peoples' Friendship University of Russia (RUDN University), Moscow, Russian Federation

²Institute of Informatics Problems, Federal Research Center Computer Science and Control of Russian Academy of Sciences, Moscow, Russian Federation

{daraselia-av,sopin-es}@rudn.ru

Abstract

In this paper we consider a multi-server model in terms of a resource loss system with the waiting buffer and the single type of resources. A customer accepted for servicing occupies a random amount of resources with described distribution functions. Based on the assumptions of a Poisson arrival process and exponential service times, we analytically find the the system of equilibrium equations, solving which we get the stationary probabilities, loss probability, the average waiting time and the average number of customers.

Keywords: queuing system, resource loss system, multi-service network, queuing system with resources, random amount of resources, arithmetic probability distribution

1. Introduction

Resource loss system (LS) has been studied for a long time. The first studies [1] explored models with the allocation of a certain random amount of resources of a single type to each arriving customer. In [2], these calculations were presented for finite LS with arbitrary distribution functions of service time and resource. Further, in [3, 4] the service time and the amount of the allocated resource were considered dependent random variables. Subsequently, this area of research has been actively developed. Interest in Resource LS can be explained by the possibility of their application for modeling a fairly wide range of technical devices and, in general, information and computing systems, wireless networks. In [5] the authors present an overview of the resource loss systems used for modeling of a wide class of real wireless networks systems with admittedly limited resources. In particular, in a number of

studies, resource LSs were used as one of the tools for modeling Long-Term Evolution (LTE) [14, 15] and 5G New Radio (NR) systems [16, 17].

Due to the complexity of the analytical calculation and modeling the corresponding random processes, most of the research results were obtained under simplifying assumptions, such as deterministic resource customers, exponentially distributed service time, the simplest incoming flow of customers, the simplest LS configuration [10]. An overview of these studies can be found in [5, 6, 7, 8, 9]. The papers [10, 11] study multiphase resource LS with a non-Poisson arriving flow, non-exponential service, an infinite number of devices, and an unlimited amount of allocated resource on each of the phases. In [13, 14] the authors study resource LS with signals, which allow redistributing resources and, in turn, takes into account the mobility of subscribers of cellular networks. However, this will not be considered in this study.

In particular, for the analysis of resource LS, there is a fundamental possibility of using well-known algorithms for the analysis of Erlang networks. In [12] the authors show the relationship between multiservice loss networks and loss systems with resources, which makes it possible to solve the problem of calculating the loss probability in the loss systems with resources using known methods developed for multiservice loss networks. However, the authors considered LS without waiting buffer where each customer only occupies a certain amount of resources. In our paper, we propose the system of equilibrium equations of the resource loss system with the waiting buffer and calculation of some technical characteristics of the system.

2. Mathematical model

Consider a multiserver resource loss system with the waiting buffer, Poisson arrivals and exponential service times. The system consists of $N < \infty$ servers and waiting buffer of size $v < \infty$, see Fig. 1. We assume the system to have a single Poisson flow of customers with arrival rate λ , and duration of customer servicing are exponentially distributed with parameter μ . Serving process of each customer requires a server and a random number r of resources, $0 \leq r \leq R$. The distribution of resource requirements are given by probability mass function (pmf) $\{p_r\}_{1 \leq r \leq R}$, where p_r is the probability that a customer requires r resources.

The system operates as follows.

- 1) An arriving customer requiring r resources is accepted for service if at the moment of arrival there is a free server and the number of occupied resources is no more than $R - r$.
- 2) An arriving customer requiring r resources takes up space in the waiting buffer if there is a room for it and at the moment of arrival there is no free server or the number of available resources is greater than $R - r$.

- 3) If there are n customers on the servers and they totally occupy r resources, then at the departure of a customer j resource units are released with probability $\frac{p_{r-j}^{(n-1)} p_j}{p_r^{(n)}}$, and one random customer from the waiting buffer is trying to enter the service.

The system behavior can be described by a random process $X(t) = (\xi(t), \delta(t)), \eta(t)$, where $\xi(t)$ is the number of customers on the servers, $\eta(t)$ is the number of customers in the waiting buffer, and $\delta(t)$ is the total amount of resources allocated at time moment t to the customers on the servers.

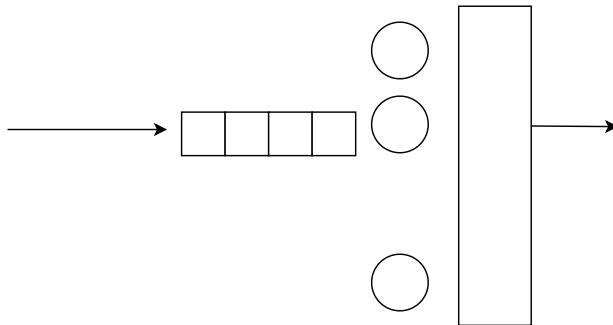


Fig. 1. Loss system with resources and waiting buffer

The state space of the system is described by the set

$$S = \bigcup_{n=0}^N S_n, \quad (1)$$

where

$$S_0 = \{(0, 0, 0)\}, \quad S_n = \left\{ (n, r, k) : p_r^{(n)} > 0, 0 \leq k \leq V \right\}, \quad 1 \leq n \leq N. \quad (2)$$

Here $\{p_r^{(n)}\}_{r \geq 0}$ is n -fold convolution of pmf $\{p_r\}_{1 \leq r \leq R}$, $p_r^{(n)}$ is the probability that n customers occupy r resource units, and is calculated as follows

$$p_j^{(n)} = \sum_{i=1}^{j-n+1} p_{j-i}^{(n-1)} p_i, \quad 1 \leq j \leq R, \quad n \geq 2, \quad (3)$$

where $p_j^{(1)} = p_j$, $p_0^{(0)} = 1$ and $p_j^{(0)} = 0$, $j \geq 0$.

Let $P_{n,k}(r)$ be the stationary probability that there are n customers on the servers that totally occupy r resources and k customers in the waiting buffer. Then the

system of equilibrium equations for this model takes the following form:

$$\lambda P_{0,0}(0) = \mu \sum_{j=1}^R P_{1,0}(j), \quad (4)$$

$$(\lambda + n\mu) P_{n,0}(r) = \lambda \sum_{j=0}^{r-1} P_{n-1,0}(j) p_{r-j} + (n+1)\mu \sum_{j=r+1}^R P_{n+1,0}(j) \frac{p_{j-r} p_r^{(n)}}{p_j^{(n+1)}} + \\ + n\mu \sum_{j=n}^R P_{n,1}(j) \sum_{i=1}^{j-n+1} \frac{p_i p_{j-i}^{(n-1)}}{p_j^{(n)}} p_{r-j+i}, \quad 1 \leq n < N, \quad (5)$$

$$(\lambda + N\mu) P_{N,0}(r) = \lambda \sum_{j=N-1}^{r-1} P_{N-1,0}(j) p_{r-j} + N\mu \sum_{j=N}^R P_{N,1}(j) \sum_{i=1}^{j-N+1} \frac{p_i p_{j-i}^{(N-1)}}{p_j^{(N)}} p_{r-j+i}, \quad (6)$$

$$(\lambda + n\mu) P_{n,k}(r) = \lambda \left(\sum_{j=n-1}^{r-1} P_{n-1,k}(j) p_{r-j} + \sum_{j=R-r+1}^R P_{n,k-1}(r) p_j \right) + \\ + (n+1)\mu \sum_{j=r+1}^R P_{n+1,k}(j) \frac{p_{j-r} p_r^{(n)}}{p_j^{(n+1)}} \sum_{i=R-r+1}^R p_i + \\ + n\mu \sum_{j=n}^R P_{n,k+1}(j) \sum_{i=1}^{j-n+1} \frac{p_i p_{j-i}^{(n-1)}}{p_j^{(n)}} p_{r-j+i}, \quad 1 \leq k < V, \quad 1 \leq n < N, \quad (7)$$

$$\left(\lambda \sum_{j=1}^{R-r} p_j + n\mu \right) P_{n,V}(r) = \lambda \left(\sum_{j=n-1}^{r-1} P_{n-1,V}(j) p_{r-j} + \sum_{j=R-r+1}^R P_{n,V-1}(r) p_j \right) + \\ + (n+1)\mu \sum_{j=r+1}^R P_{n+1,V}(j) \frac{p_{j-r} p_r^{(n)}}{p_j^{(n+1)}} \sum_{i=R-r+1}^R p_i, \quad 1 \leq n < N, \quad (8)$$

$$(\lambda + N\mu) P_{N,k}(r) = \lambda \sum_{j=N-1}^{r-1} P_{N-1,k}(j) p_{r-j} + \lambda P_{N,k-1}(r) + \\ + N\mu \sum_{j=N}^R P_{N,k+1}(j) \sum_{i=1}^{j-N+1} \frac{p_i p_{j-i}^{(N-1)}}{p_j^{(N)}} p_{r-j+i}, \quad 1 \leq k < V, \quad (9)$$

$$N\mu P_{N,V}(r) = \lambda \sum_{j=0}^r P_{N-1,V}(j)p_{r-j} + \lambda P_{N,V-1}(r), \quad (10)$$

where $1 \leq r \leq R$.

The system of equations (4)-(10) for stationary probabilities $P_{n,k}(r)$, $0 \leq k \leq V$, $0 \leq n \leq N$, $0 \leq r \leq R$ is solved numerically.

With the stationary distribution, one can obtain the loss probability π .

$$\pi = \sum_{r=N}^R P_{N,V}(r) + \sum_{n=1}^{N-1} \sum_{r=n}^R P_{n,V}(r) \sum_{i=R-r+1}^R p_i. \quad (11)$$

The average number \bar{N} of customers in the waiting buffer is

$$\bar{N} = \sum_{n=1}^N \sum_{r=n}^R \sum_{k=1}^V k P_{n,k}(r). \quad (12)$$

The average waiting time W is determined by well-known Little's law:

$$W = \frac{\bar{N}}{\lambda(1 - \pi)}. \quad (13)$$

3. Conclusion

In our paper we consider a model of a multi-server resource loss system with the waiting buffer and derived the system of equilibrium equations for the stationary probabilities of the system. Besides, we obtained the loss probability of the system, the average waiting time and the average number of customers.

4. Acknowledgement

The research was funded by the Russian Science Foundation, project no.20-07-01052.

REFERENCES

1. Romm E, L., Skitovitch V. V. On certain generalization of problem of Erlang // Automation and Remote Control. 1971. V. 32 No. 6. P. 1000—1003.
2. Tikhonenko O.M., Determination of the Characteristics of Queueing Systems with Limited Memory // Autom. Remote Control, 1997, vol. 58, no. 6, pp. 969–973.

3. Tikhonenko O.M. and Klimovich K.G., Analysis of Queueing Systems for Random-Length Arrivals with Limited Cumulative Volume // Probl. Peredachi Inform., 2001, vol. 37, no. 1, pp. 78–88.
4. Tikhonenko O.M., Generalized Erlang Problem for Queueing Systems with Bounded Total Size // Probl. Peredachi Inform., 2005, vol. 41, no. 3, pp. 64–75.
5. Gorbunova A. V. et al. Resource queuing systems as models of wireless communication systems // Informatics and Applications. 2018. V. 12 No. 3. P. 48–55.
6. Naumov V. et al. On the total amount of resources occupied by serviced customers // Automation and Remote Control. 2016. V. 77. P. 1419–1427.
7. Basharin G.P. et al. A new stage in mathematical teletraffic theory // Automation and Remote Control. 2009. V. 70. No. 12. P.1954–1964.
8. Naumov V., Samouylov K. Analysis of multi-resource loss system with state-dependent arrival and service rates // Probability in the Engineering and Informational Sciences. 2017. V. 31. No. 4. P.413–419.
9. Gorbunova A. V. et al. Resource queuing systems with general service discipline // Inform. Primen., 13:1 (2019), 99–107
10. Galileyskaya A.A. et al. On Sequential Data Processing Model, that Implements the Backup Storage // Modern Information Technologies and IT-Education. 2019; 15(3):579–587.
11. Lisovskaya E. et al. Infinite-server tandem queue with renewal arrivals and random capacity of customers // Comm. Com. Inf. Sc., 2017. Vol. 700. P. 201–216.
12. Naumov V.A. and Samouylov K. E. On relationship between queuing systems with resources and Erlang networks // Informatics and Applications. 2016. V. 10 No. 3. P. 9–14.
13. Ageev K. et al. Simulation of the Limited Resources Queueing System with Signals // 2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Moscow, Russia, 2018, pp. 1–5.
14. Sopin E. et al. LTE network model with signals and random resource requirements 2017 // 9th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). 2017. P. 101–106.
15. Sopin E.S. et al. Performance Analysis of M2M Traffic in LTE Network Using Queuing Systems with Random Resource Requirements // Aut. Control Comp. Sci. 2018. V. 52. P.345–353.
16. Lu X. et al. Integrated Use of Licensed- and Unlicensed-Band mmWave Radio Technology in 5G and Beyond // IEEE Access. 2019. V. 7. P. 24376–24391.
17. Daraseliya A. V. et al. Analysis of 5G NR Base stations offloading by means of NR-U technology // Informatics and Applications. 2021. V. 15, No. 3. P. 98–111.

UDC: 519.25

The error correction method in the problem of automatic authorship identification of literary text

Yu. N. Orlov¹ and M. Yu. Voronina¹

¹Keldysh Institute of Applied Mathematics, Miusskaya sq., 4, Moscow, 125047, Russia
ov3159f@yandex.ru, voronina.miu@phystech.edu

Abstract

The paper describes an algorithm for correcting of possible identification errors of authors of literary texts. The identification method is the nearest neighbor pattern, corresponding to a given class of texts. The pattern in this case is the empirical frequency distribution of letter combinations based on the analysis of reliably known works of the author. The proximity between texts is understood in the sense of the proximity of the frequencies of bigrams in the L1 norm. The author of an unknown text is assigned the one to whose pattern the text under test is closest. For identification, a library of authors is used, each of which has a sufficiently large number of works defining the corresponding patterns of two-letter combinations. In the analyzed corpus of texts, 1783 texts of 100 authors were collected, the recognition error was equal to 0.12. It is important that after the exclusion of incorrectly recognized texts, a library of 88 authors and 1,450 texts remained, each of which was identified correctly. The problem under study is the assessment of the probability that there is no standard of the author of the tested text among the library patterns. To solve it, the paper analyzes the dependence of the probability of erroneous identification on the length of the text. Using the example of an unmistakably determined subgroup of texts, it turned out that the empirical probability of correct recognition of a text fragment, although it decreases with a decrease in the length of the fragment, still exceeds 0.5 up to the fragmentation of the text into 10 parts. If the correct pattern is excluded from consideration, the second closest pattern is assigned as such, but it turns out to be unstable: the ambiguity of such identification of the author of fragments occurs already when the text is cut into 4 fragments. Thus, the stability of the identification of the author of text fragments can be proposed as an independent criterion for the correctness of the method that allows you to select texts written in atypical styles.

Keywords: text, author, frequency distribution of letter combinations, fragment identification, error correction

1. Introduction

The task of automatic classification of text by attributes is very relevant in the context of the development of information technology and big data analysis. In addition, it is located at the intersection of sciences: mathematical statistics and linguistics, which can contribute to the development of both branches of knowledge. However, it should be noted that the interpenetration of the methods used in these sciences is very slow. This is partly due to the fact that numerous machine learning methods (see e.g. [1, 2, 3, 4]), used for various processing and analysis of text documents represent a "black box" not only for linguists, but also for mathematicians themselves. It is not clear to what extent these methods, configured to solve the classification problem within a specifically selected corpus of texts, are robust in relation to other samples and other tasks. In addition, due to the specifics of the setting, it is difficult to propose a procedure for its correction.

In the overview paper [5] the effectiveness of purely statistical methods of analysis based on the calculation of formal indicators, such as the number of letters, words, punctuation marks, etc., is compared with expert methods of analyzing the author's style, turns of speech, the use of literary techniques. The authors conclude that although the expert method is more valuable for literary critics, it does not have sufficient accuracy on a large body of texts and, more importantly, cannot be adequately implemented in the form of a formal computer program. At the same time, the statistics of letters or letter combinations, although they do not have a direct literary meaning, can be quite unambiguously compared to each text with an indication of the error within the formal criteria. Thus, in the context of the task of machine recognition of text attributes, the statistical method is more effective, i.e. it has less error than the expert method.

In identification tasks, a significant problem after the recognition model itself is the assessment of the probability of making an erroneous decision: either a false acceptance of "someone else's" for "one's own", or a false rejection of "one's own" taken for "someone else's". Formally recognized object in mathematical terms is a numerical vector, the dimension of which is equal to the dimension of the space of parameters measured during the observation of the object. For example, if the frequency of occurrence of letters of the Russian alphabet in any texts is studied, then such a vector has dimension 33, bigram vector has dimension $1089 = 33^2$ etc.

The idea of statistical recognition is that each object is perceived not exactly, but with some error generated by the finite accuracy of measurements and the finite length of the studied data series. Then the result of a separate observation is a sample distribution from the general set of parameters that presumably correspond to the ideal model. For example, if a number of human biometric data is observed, such as body temperature and heart rate, then without taking into account changing

external conditions, not the unique temperature value will be obtained, but some of its distribution, and the distribution of frequency values is similar. These distributions will represent a separate empirical sample for a certain period of time. The totality of such samples over a long period of observations form a pattern. With this approach, the coordinates of the observed object lie in some multidimensional parallelepiped, the length of the side of which for each variable corresponds to the spread of data near the reference value. An obvious sufficient condition for the correct identification of an object is the non-overlap of the carriers of parameter distributions for different objects. However, in reality, this situation is very rare. Therefore, the belonging of the observation result to one or another set is determined, generally speaking, with some non-zero error. The probability of an error depends on the comparison rule and the metric in which objects are compared with patterns.

In practice, standards are more or less well-founded theoretical models or are created based on the results of processing a large array of empirical data. The technology of artificial intelligence implementing the recognition task will be called logically transparent if an unambiguous identification method is formulated, each step of which can be verified, and the probability of error is estimated by explicitly formulated criteria outside the machine learning system with a list of a priori assumptions about the properties of objects of the studied category. This definition clarifies the part of the error assessment the agreement adopted in the work [6] that explicable (interpreted, transparent) artificial intelligence is a machine learning method that can be called a "white box", and which allows you to check exactly which operations and at what stages were carried out by the recognizing algorithm.

Statistical identification consists in finding the most probable general population to which this sample distribution could belong. The error of such a decision is given by the Kolmogorov criterion. If the sample is large enough, as is the case with the analysis of large literary works, then the confidence level is close to unity. However, it turns out that the general aggregates characterizing the authors (i.e., the author's patterns) differ very little from one another, so that the formal proximity to someone else's pattern also turns out to be almost the same as for their own pattern. In this regard, there is a need to develop a second indicator, in addition to the actual distance between the sample and the pattern, which would give an estimate of the probability that the nearest neighbor found among the author's patterns is not the correct answer. The difficulty of constructing such an indicator is that, on the one hand, it should not depend on the first indicator, i.e. on the proximity of the distributions of letter combinations of the text and the pattern. But on the other hand, in the case of independence of indicators, the probability of correct recognition by two indicators is less than by each of them separately. If the

indicators are dependent, then there is no need for such a composition, because then the best option is to use the most accurate indicator.

This paper presents a model for correcting the identification error using the example of the problem of recognizing the author of a text by the method of cross-validation. The author's pattern is the empirical frequency distribution of letter combinations, constructed according to all reliably known works of the author. The proximity between texts is understood as the proximity between the empirical frequency distributions of paired letter combinations – bigrams in the sense of the norm in L_1 . The author of an unknown text is assigned the one to whose pattern the text under test is closest, i.e. the nearest neighbor method is used. For identification, a library of authors is used, each of which has a sufficiently large number of works defining the corresponding patterns of bigrams. The problem under study is the assessment of the probability that there is no pattern of the author of the tested text among the library patterns. To solve it, it is proposed to investigate the dependence of the probability of erroneous identification on the length of the text. Numerical experiments have shown that the correct author is consistently recognized on fragments of text significantly (by an order of magnitude) smaller than the original one. If the correct pattern is excluded from consideration, the identification will be obviously incorrect, and the "author" will be answered by the next closest standard, which, when fragmenting the text, may show properties different from the correct standard of the author. The stability of the identification of the author of text fragments is proposed as a new criterion for the correctness of the method.

The frequencies of bigrams for recognizing the authors of texts have been used before. In the works [7, 8] the text was considered as the trajectory of the Markov process of joining letters, taking into account certain grammatical rules, for which it was necessary to estimate the conditional probability of the appearance of characters in the text. In the monograph [9] various metrics were studied to determine different attributes of the text, also based on empirical frequencies of letter combinations. However, in the mentioned works, the analysis was carried out on a relatively small corpus of texts (about several dozen authors).

Note that there are a large number of semantic identification methods (see, e.g. the overview [10]), based on the analysis of the frequency of use of words. Each of these methods has an independent field of application, but in general these are "engineering" techniques, the reliability of which does not exceed 0.7 for a sufficiently large body of texts. We do not aim here to compare different methods, but demonstrate the effectiveness of the bigram method and the associated text fragmentation method, followed by "voting" on the frequency of mentioning the author of the fragment.

2. The construction of author patterns

The concept of recognizing the author of a text by empirical frequencies of letter combinations is based on the assumption that each professional writer corresponds to his personal pattern, interpreted as a general set. Then individual texts are samples from this totality and deviate from it due to the finiteness of the text and some differences in subject matter between the texts. Let $F_a(j)$ is a frequency of symbol j in a pattern of author a , where j designates a corresponding bigram. Symbols of other alphabets, spaces, punctuation marks and numbers are ignored. Uppercase and lowercase letters do not differ. Let also $D_a^i(j)$ is an empirical frequency of symbol j in the i -th text of author a and N_a^i is a number of symbols in this text. Let n_a is a number of texts of author a . Then the empirical estimation of pattern $F_a(j)$ of the author a is constructed as a weighted distribution over the total set of the texts of a given author:

$$F_a(j) = \frac{1}{N_a} \sum_{i=1}^{n_a} N_a^i D_a^i(j), \quad N_a = \sum_{i=1}^{n_a} N_a^i. \quad (1)$$

The distance between the i -th text of author a and a pattern $F'_a(j)$ of the same author is define by formula:

$$x_{aa}^i = \sum_{j=1}^J |D_a^i(j) - F'_a(j)|, \quad (2)$$

where $J = 33^2$ and a streak $F'_a(j)$ means, that this text is excluded from the pattern of corresponding author (1):

$$\begin{aligned} F'_a(j) &= \frac{1}{N_a - N_a^i} \sum_{\substack{k=1 \\ k \neq i}}^{n_a} N_a^k D_a^k(j) = \\ &= \frac{1}{N_a - N_a^i} \sum_{k=1}^{n_a} N_a^k D_a^k(j) - \frac{N_a^i D_a^i(j)}{N_a - N_a^i} = \\ &= \frac{N_a F_a(j) - N_a^i D_a^i(j)}{N_a - N_a^i}. \end{aligned}$$

Then the distance from i -th text to the pattern of the real author has the form:

$$\begin{aligned} x_{aa}^i &= \sum_{j=1}^J \left| D_a^i(j) - \frac{N_a F_a(j) - N_a^i D_a^i(j)}{N_a - N_a^i} \right| = \\ &= \frac{1}{1 - N_a^i/N_a} \sum_{j=1}^J |D_a^i(j) - F_a(j)|. \end{aligned} \quad (3)$$

It should be noted that these formulas refer to the text that is considered correct, that is, written without errors and typos. Each typo distorts the distribution of bigrams by an amount not exceeding $\delta = 4/N$, where N is a text length. Since the average length of the text is about 400 thousand characters, then k typos may lead to a distortion of the distribution of bigrams by $\varepsilon = 10^{-5}k$. In addition, the books inevitably contain words that are not related to the author's style: these are epigraphs, as well as the numbering of parts and chapters and their names. These aspects were not taken into account in the analysis. Therefore, it is necessary to estimate the order of error that is allowed with such an approximation.

The empirical frequency of bigram distribution for descending order of frequency is presented in Fig. 1.

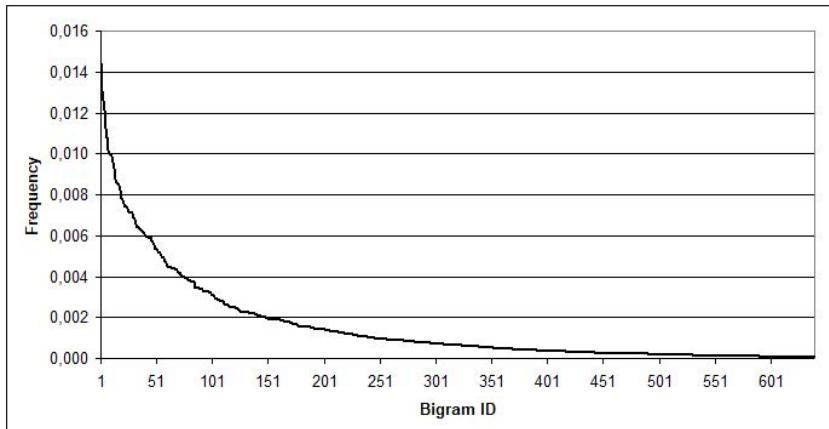


Fig. 1. The bigram distribution for descending order of frequency

The typical values of the distances (3) lie from 0.1 to 0.2. The distribution of these distances over the corpus under testing is presented in Fig. 2.

The distribution of distance difference between text and the first and the second patterns is presented in Fig. 3.

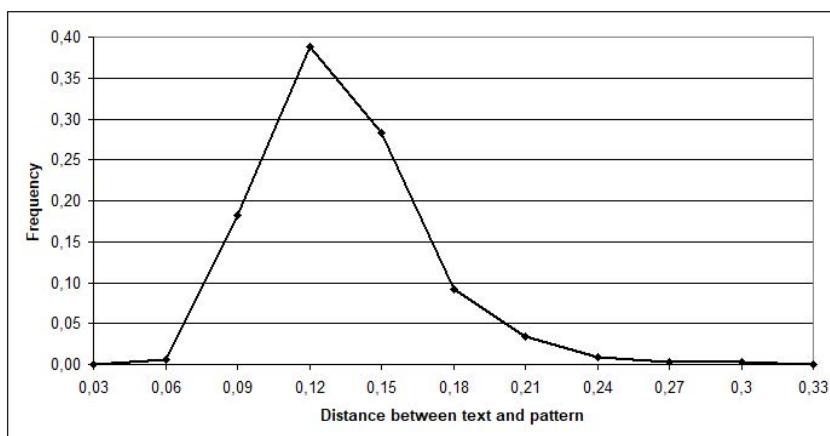


Fig. 2. The distance distribution “text-the real author pattern”

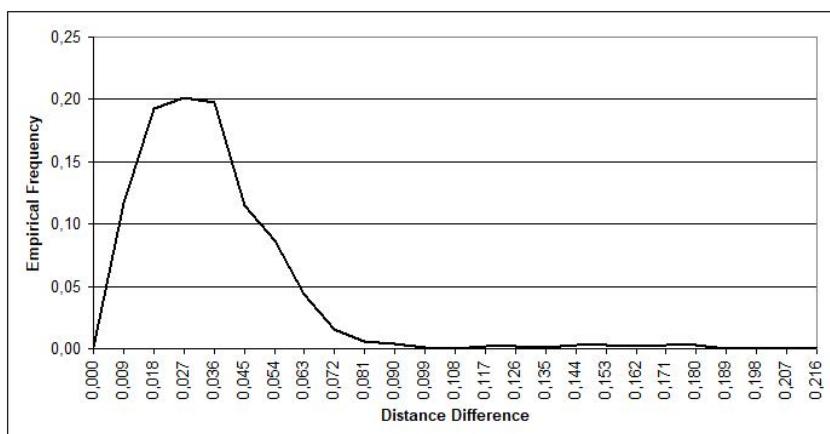


Fig. 3. The distribution of distance difference between text and the first and the second patterns

From Fig. 3 it follows that the main part of the distribution of the distance difference from the text to the correct and to the nearest incorrect patterns is contained in the interval from 0.01 to 0.06. The probability that the specified difference is less than 0.01 is approximately 0.12. Therefore, in order to significantly affect the recognition result, the distortion of the bigram distribution should be on the order of 0.01, which corresponds to about 1000 typos per book or 3 typos per page. This is a very large amount that is not found in modern printed products.

Therefore, there is no need for pre-processing of texts in order to eliminate possible typos.

The distance between i -th text of the author a and the pattern $F_b(j)$ some other author $b \neq a$ is defined as

$$y_{ab}^i = \sum_{j=1}^J |D_a^i(j) - F_b(j)|. \quad (4)$$

So the final formula for the distance between text and pattern has the form:

$$z_{ab}^i = \frac{1}{1 - \delta_{ab} N_b^i / N_b} \sum_{j=1}^J |D_a^i(j) - F_b(j)|, \quad (5)$$

where δ_{ab} is the Kronecker symbol.

So the author b of any unknown text with the bigram distribution $D(j)$ is determined through the condition:

$$z(a) = \sum_{j=1}^J |F_a(j) - D(j)| = \min \Rightarrow b = \operatorname{argmin} z(a) \quad (6)$$

However, if there is no pattern of the author of the text under study, the nearest pattern will still be found according to rule (6) but, of course, the resulting answer will be incorrect. In this paper, it is proposed to use a second indicator based on the hypothesis of the stability of "correct" pattern and the instability of "wrong" pattern when dividing the text into fragments.

On the formal side, the problem boils down to estimating the probability of correct identification of samples from different general distributions, depending on the proximity between these distributions and the sample length. As far as the authors of this paper know, no large-scale statistical experiment of this kind has been conducted for some standard distributions. If the distributions of letter combinations were stationary, the error estimate would be the highest level of significance at which the sample distribution is recognized in accordance with the Kolmogorov criterion. At the same time, the ordering of the patterns of "wrong" authors would be preserved by the distance to the texts of this author in accordance with which patterns are close to the pattern of this author. However, as studies in the field of statistics of texts in natural languages show, the frequency distributions of letter combinations are non-stationary. It is this aspect that allows you to determine the error of recognizing the author of the text in the absence of one in the library of patterns.

Let's test a sufficiently long text that can be "cut" into a certain number of fragments. By reducing the length of a text fragment, we thereby increase the

statistical uncertainty of estimating sample frequencies. At the same time, it was noticed that the patterns of "wrong" authors, following the correct standard in the order of increasing the distance to the full text, do not have a stable ordering: with a decrease in the length of the fragment, the patterns of different authors are in second place. Therefore, if you remove the pattern of "correct" author from the library, then the patterns closest to the text will not have the properties of stable identification of the text with a decrease in the length of the fragment, whereas the correct pattern has the stability, mentioned above. Then, by associating with each author the probability distribution of erroneous identification of a text fragment depending on its length, we obtain another set of distributions with which the empirically obtained probabilities of erroneous identification will be compared, assuming that the primary identification was carried out correctly.

In relation to the problem under study, the minimum length of the fragment is chosen equal to 1000 symbols. Although this smallness is redundant, since such samples are not representative, it is adopted for greater completeness of the analysis and for the convenience of variation of the sample length by small fragments of text. For text with N symbols we have $n = [N/1000] + 1$ fragments. The first $n - 1$ of fragments have a length 1000 symbols, and the length of the last fragment is equal to $l = N - 1000[N/1000]$. If the length of the last fragment is less than half of the minimum (i.e. less than 500 characters), then this fragment is combined with the penultimate one. If it is more than 500 characters, then the fragment is analyzed in the same way as the rest.

To save calculations, the library is read only once – for minimal fragments of texts. For each text we construct bigram distributions $f_k^{(1)}(j)$, $k = 1, 2, \dots, n - 1$ of the fragments with 1000 symbols. Hence we can construct the distributions $f_k^{(2)}(j)$ of the fragments with 2000 symbols etc.

If we have the fragment with $s \cdot 1000$ symbols, the corresponding bigram distributions has the form:

$$f_k^{(s)}(j) = \frac{1}{s} \left(f_{s(k-1)+1}^{(1)}(j) + f_{s(k-1)+2}^{(1)}(j) + \dots + f_{sk}^{(1)}(j) \right), \quad (7)$$

$$k = 1, 2, \dots, [(n - 1)/s].$$

The distribution $D(j)$ of the initial text is obtained from (7) by formula:

$$D(j) = \frac{1000}{N} \sum_{k=1}^{n-1} f_k^{(1)}(j) + \left(1 - (n - 1) \frac{1000}{N} \right) f_n^{(1)}(j) \quad (8)$$

After that we can construct the author patterns (1) from the distributions (8).

3. Correction of the text author recognition error

We consider the corpus of literary texts in Russian, which presents one hundred domestic and foreign (translated) authors with at least six works of at least 30 thousand characters in length. Of these authors, 40 were Russian, and 60 were foreign. The total number of texts reviewed amounted to 1,783 works with a total volume of approximately 700 million characters. The author's patterns of bigrams and the distribution of distances between the distributions of symbols in texts and patterns were constructed according to the methodology described above. It turned out that the probability of erroneous identification of the author of a separate text was 0.12.

Fig. 4 shows the joint distribution of "true-wrong" distances for this case. The axes of distances from the text to the authors' pattern are divided into 20 class intervals in increments 0.03. The surface presented in this figure PROB(TRUE, WRONG) shows the part of the texts, the distance from which to correct pattern fell into the specified cell TRUE, and to incorrect pattern fell into the specified cell WRONG. Since almost the entire carrier of this distribution is located to the right of the side diagonal of the square, this means that almost always the distance to one's own standard is less than to someone else's.

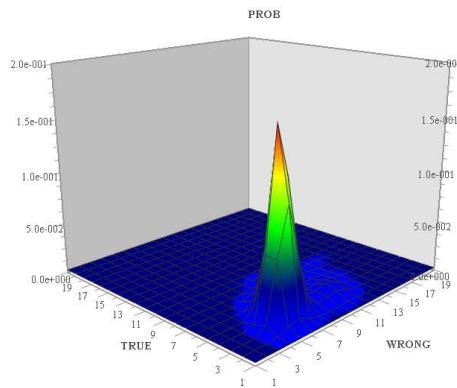


Fig. 4. Joint density of the distribution of distances of bigrams between texts and patterns of "true-wrong"

Next, we will make one important remark related to the adequacy of the applied hypothesis of recognition of the author. This method allows for the so-called "cleaning selection". It consists of the following. At first, all mistakenly identified texts were excluded from the library, and if as a result some author had less than six works, then he was completely excluded from this corpus. After that, new patterns were

compiled, and identification was carried out again. Of course, since the patterns changed, texts appeared again whose authors were incorrectly identified. Such texts were again excluded. It is important to emphasize that after the six iterations described, 88 authors (38 Russian and 50 foreign) and 1452 texts remained in the corpus, the author of each of which was identified unmistakably. The main statistical experiment on the identification of text fragments depending on their length was carried out on this cleaned corpus. Note that in many statistical models that are not based on the analysis of the content of the studied systems, the gradual elimination of erroneous results leads to the almost complete exclusion of sample elements: such, for example, regression models.

In a more general case, a joint distribution $f^a(\tau, \nu)$ is constructed for the author a . This distribution is constructed according to the calculated data $\nu_i^a(\tau)$, corresponding to the texts of this author a . The interval $\tau \in [0; 1]$ is divided onto L classes, after that it is determined what part of the works of author a fell into the given cell carrier $[(n-1)/L; n/L] \times [(k-1)/h; k/h]$ of the distribution $f^a(\tau, \nu)$. This share is actually an empirical estimate of this distribution in the form of $f^a(n/L, k/h)$. In Fig. 5 we show an example of such a distribution for one of the authors of the text corpus. It can be seen that the error is quite small already on the fractions of the text of the order of 0.01, that is, by 3-4 thousand characters. However, it should be noted that this approach is suitable only for authors who have many works, at least several dozen. Otherwise, such a distribution turns out to be unrepresentative.

According to the texts from the cleaned corpus the following sequence of fragments is determined $\{\tau_0, \tau_0 + 1, \dots, \tau_0 + s\}$, for which under the every value of τ the part of correctly identified fragments $\nu_i^a(\tau)$ is non-zero. For example, it turned out that if you divide the full text into two equal fragments, then with a probability of about 0.95 the author of the full text is also identified as the author of both fragments, but there are examples when he is not recognized for one or both fragments. At the same time, at each level of fragmentation from the full text, the correct author is always present as the author of at least one fragment. This empirical fact can be used to correct the result of full text recognition. As an illustration of the idea, the diagram in Fig. 6 shows the empirical frequencies with which foreign authors meet as authors of at least one fragment at the intersection of responses for all three levels of fragmentation.

It can be seen from Fig. 6 that the closer the pattern of someone else's author is to the text, the greater the probability of erroneous identification of this particular author. Although this result seems quite natural, it indirectly confirms the adequacy of the author's by no means obvious concept of recognition by proximity to the reference distribution by characters (not by words!) as to the general population.

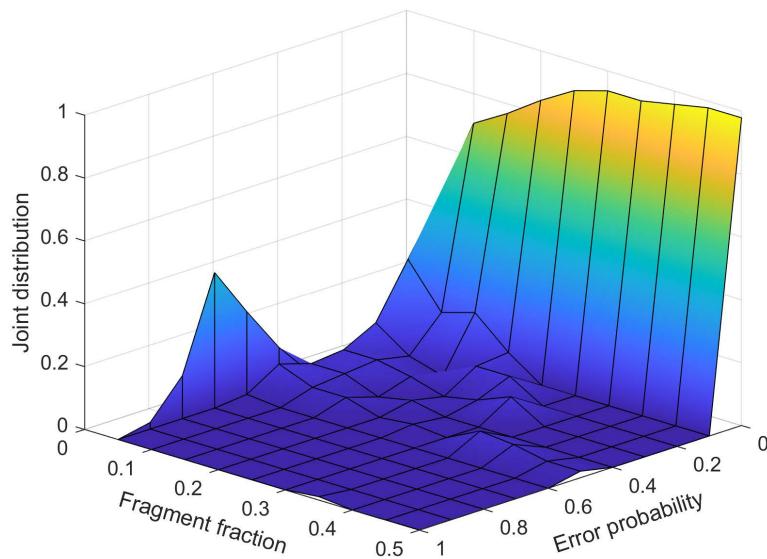


Fig. 5. Joint distribution of error probability and fragment fraction

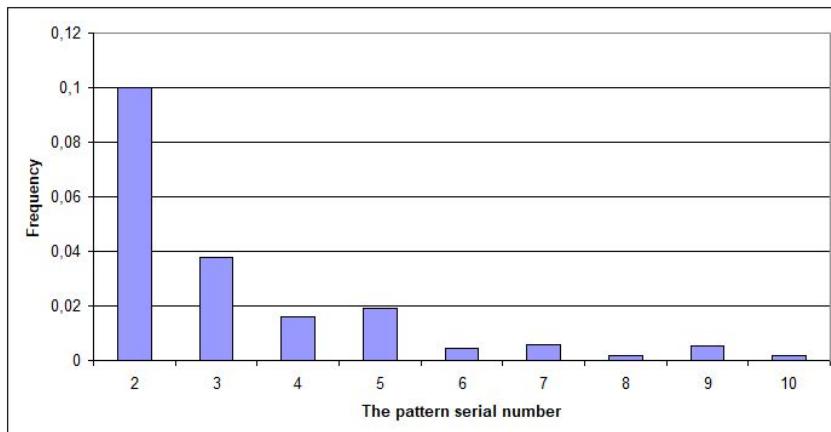


Fig. 6. The frequency of occurrence of someone else's author as the author of a fragment at the intersection of three levels of fragmentation, depending on the reference number of this author in relation to the full text

Note now that for incorrectly identified full texts, the author who is assigned first is not present in 90% of fragmented texts, as the correct author in the cleaned corpus,

but only in 47%. Thus, with respect to about half of the erroneous identification cases, it will be possible to make a corrective conclusion that the recognition is not recognized as correct.

Therefore, this correction method can also be used as an indicator that the author is presumably absent from this library of standards. Let's assume that some text was recognized as the author's *a* text. But if we mentally exclude this author from this corpus, then the author of the text in question will be assigned the next author *b*, i.e. the second closest standard to this text. Then dividing the studied text into fragments of length $\{\tau_0, \tau_0 + 1, \dots, \tau_0 + s\}$, one can obtain the parts $\nu^b(\tau_0), \nu^b(\tau_0 + 1)$ etc., corresponding to the identification of these fragments as texts of author *b*. If any one of these values turns out to be zero, then we will conclude that the original recognition of the full text was erroneous.

In general, according to the corpus of "purified" texts, its author quite often turns out to be the most identifiable author of fragments. However, it should not be assumed that fragmentation can replace the full text in accuracy.

Fig. 7 shows the dependence of the author recognition error on the most frequent author of fragments, depending on the length of the fragment. It can be seen that with a decrease in the length of the fragment, the error increases, that is, there are texts whose fragments are more often recognized by other people's patterns. Even when dividing the text in half, there was one such example when none of the parts is recognized by its author.

Thus, if we provide the recognition result with the criterion "accept as correct" or "reject as incorrect" according to the confirmation of the second indicator described above, then all correctly identified texts will remain such, and among the rest the texts whose authors are identified incorrectly will be correctly identified. Thus, the share of erroneous decision-making will decrease to about 7% instead of the original 12%. Consequently, the proposed filter, based on the fragmentation of the source text, allows approximately halving the error of incorrect recognition of the author of the text. Similarly, fragmentation can be used to test the hypothesis that the author is missing from the library of standards. The error in this case turns out to be equal to 8%, which is close to the corrected recognition error of the present author.

4. Conclusion

The paper proposes a new method for correcting the results of machine identification of the author of the text. This method uses, in a certain sense, the idea of multiplying samples, which allows us to build a corrective indicator without compromising the accuracy of the main indicator. The correction problem is related to the fact that the result of applying the main indicator is not statistics, but a single number, which is further interpreted within the framework of the nearest neighbor

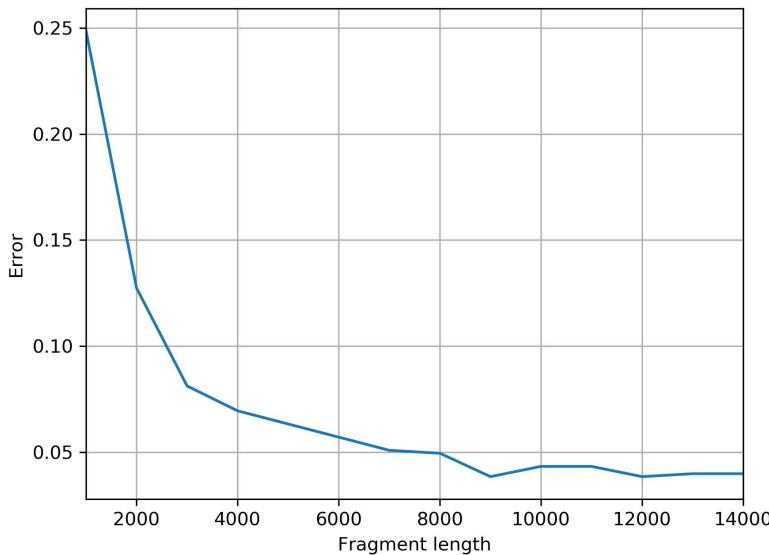


Fig. 7. Error identifying the author of the text by the most frequent author of fragments for correctly recognized full text

method. Text fragmentation allows us to get statistics of potential authors. And although fragmentation itself worsens the accuracy of recognition, it nevertheless allows us to use a new indicator as an indicator – namely, the stability of its author and the instability of someone else's. Due to the approximately uniform randomness of recognition errors, which are associated with standards far enough from the correct author (sixth and further), errors of this kind can be eliminated by fragmentation.

This method also allows us to estimate the probability that the standard of the correct author is missing from the library. The accuracy of this hypothesis is about the same as the accuracy of the method of identifying the correct author, i.e. 7%. A noticeable decrease in error on a sufficiently large corpus of texts shows that this method is very effective. At the same time, we do not discuss the reasons why the author who is actually present in the library is not correctly identified. However, preliminary analysis shows that most of these incorrectly recognized texts are located far from all authors, not just from their own, and only by chance one of them turned out to be the nearest neighbor. It can also be understood that the author wrote the text in a manner different from his other works, and in this sense he acts as some new author, whose pattern is really missing in the library. The issue of clustering texts

within the works of the same author is a topic of special research, but already now we can say that in some cases such clustering really leads to improved recognition. In particular, this applies to the so-called serial writers who produce, for example, several science fiction novels in the genre of Star Wars, and then write a number of historical novels or fantasy. That is why we interpret the correct deviation of the result of erroneous identification as a correction. Further research will allow us to formalize this approach in terms of distances to the reference two-dimensional distributions of the authors.

In conclusion, we also note that the developed method can be applied to similar tasks of identifying sample distributions in other subject areas.

Acknowledgements: This publication has been prepared with the support of Russian Ministry of Science and Higher Education, agreement № 075-15-2020-808.

REFERENCES

1. Stamatatos E., Fakotakis N., and Kokkinakis G. Automatic Text Categorization in Terms of Genre and Author // Computational Linguistics. 2000. V. 26(4). P. 471-495
2. Argamon S. and Juola P. Overview of the International Authorship Identification Competition at PAN-2011. // Petras V., Forner P., Clough P. D. (eds.) CLEF Notebook Papers/Labs/Workshop. 2011.
3. Sudheep E. M., Chinchu J., Puthussery A. and Sasi N. K. Text classification for authorship attribution analysis // Advanced Computing: An International Journal (ACIJ). 2013. V.4. No.5
4. Cappellato L., Ferro N., Halvey M., and Kraaij W. (eds.). CLEF 2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org). 2014
5. Rezanova Z. I., Romanov A. S., Mescheriakov R. V. O vbyore priznakov teksta, relevantnyh v avtorovedcheskoi i ekspertnoi dejatelnosti (in Russian) // Bulletin of Tomsk State University. Philology. 2013. V. 26. No. 6. P. 38–52.
6. Roscher R., Bohn B., Duarte M. F., Garcke J. Explainable Mashine Learning for Scientific Insights and Discoveries // IEEE Access. 8, 2020. doi: 10.1109/ACCESS.2020.2976199
7. Khmeliov D. V. Raspoznavanie avtora texta s ispolzovaniem tsepei Markova (in Russian) // Vestnik MGU, Ser. 9: Filologija. 2000. V. 2. P. 115–126.
8. Kukushkina O. V., Polikarpov A. A., Khmeliov D. V. Raspoznavanie avtorstva texta s ispolzovaniem bukvennoi I grammaticeskoi informatsii (in Russian) // Problemy peredachi informatsii. 2001. V. 37. No 2. P. 96–109

9. Orlov Yu. N., Osminin K. P. Metody statisticheskogo analiza literaturnykh tekstov (in Russian). Editorial URSS. 2012.
10. Batura T. V. Metody avtomaticheskoi klassifikatsii textov (in Russian) // Programmnye produkty i sistemy. 2017. V. 30. No. 1. P. 85–99.

УДК: 519.218.31

О периоде занятости и загрузке системы обслуживания с разделением времени в случайной среде

А.В. Зорин¹

¹ННГУ им. Н.И. Лобачевского, проспект Гагарина, 23, Нижний Новгород, Россия

andrej.zorine@itmm.unn.ru

Аннотация

В работе рассматривается определение понятия загрузки как предельного среднего времени активности прибора. Приводятся расчетные формулы для определения загрузки однолинейной системы массового обслуживания с повторными вызовами и входным потоком, формируемым в случайной среде с двумя состояниями. От состояния среды зависят интенсивность поступления групп требований и распределение размера группы. Такая модель описывает динамику очереди в системах обработки однородной информации с разделением времени. Предполагается наличие стационарного режима функционирования СМО. Необходимые и достаточные условия существования стационарного режима также найдены.

Ключевые слова: Входные потоки в случайной среде, обслуживание с разделением времени, загрузка системы массового обслуживания

1. Введение

Для классической системы $M/M/1/\infty$ с интенсивностью входного потока λ и интенсивностью обслуживания μ величину $\rho = \lambda/\mu$ часто называют *загрузкой* системы. То есть, имеет место ситуация, когда указывается, *как* вычислять, но нет определения, *что* вычисляется, т.е., что значит термин “загрузка” в этом контексте. Если исходить из того, что в стационарном режиме ρ есть вероятность непустой очереди, было бы верней называть загрузкой величину $\min\{1, \rho\}$, поскольку при отсутствии стационарного режима загрузка будет уже равна 1, а не ρ . Поэтому совершенно справедливо ряд авторов называют величину ρ ожидаемым количеством работы, поступающим в единицу времени.

В рамках данной статьи мы будем следовать определению загрузки системы как доли времени, в течение которого обслуживающее устройство занято работой, и приведем алгоритм вычисления этой доли времени для однолинейной системы

обслуживания с входным потоком с переменной структурой и с повторным обслуживанием требований (известным также как *бернульиевская обратная связь*). Системы такого типа являются моделями процессов обработки информации в системах с разделением времени однородных требований. Для этого придется также рассмотреть период занятости указанной СМО.

2. Описание системы обслуживания

Рассмотрим однолинейную систему массового обслуживания, в которую поступает входной поток требований Π , формируемый внешней случайной средой с двумя состояниями, $e^{(1)}$ и $e^{(2)}$. Заявки этого потока будем называть первичными. На интервале пребывания среды в состоянии $e^{(k)}$ требования по потоку Π образуют пуссоновский поток групп с параметром $\lambda^{(k)}$, а распределение числа заявок в группе имеет производящую функцию $f^{(k)}(v) = \sum_{w=1}^{\infty} f_w^{(k)} v^w$, аналитическую в круге $|v| < 1 + \varepsilon$ для некоторого $\varepsilon > 0$. Здесь величина $f_w^{(k)}$ определяет вероятность того, что число заявок в группе равно w , когда состояние случайной среды есть $e^{(k)}$. Для требований имеется накопитель O_1 с неограниченной очередью. В каждый момент времени обслуживается не более одного требования и это обслуживание совершается без прерывания. Если по окончании обслуживания очередь не пуста, то на обслуживание выбирается новое требование из очереди, для определенности — из головы очереди. Если же заявок в очереди нет, то прибор начинает обслуживать первое пришедшее первичное требование. Длительность обслуживания произвольного требования задается функцией распределения $B(t)$, $B(+0) = 0$. Длительности обслуживания требований предполагаются независимыми между собой и независимыми от входного потока требований.

Математической моделью функционирования внешней среды в простейшем случае является однородная неприводимая непериодическая марковская цепь с двумя состояниями $e^{(1)}$ и $e^{(2)}$. Смена состояний марковской цепи может происходить только в моменты окончания актов обслуживания. Задана вероятность $a_{k,l}$ перехода этой цепи из состояния $e^{(k)}$ в состояние $e^{(l)}$ за один шаг, $k, l \in \{1, 2\}$.

После обслуживания требование может с вероятностью p , $0 \leq p < 1$, поступить в очередь O_1 на повторное обслуживание, образуя поток вторичных требований, а с вероятностью $1 - p$ может покинуть систему.

Пусть $\tau_0 = 0$ и τ_i , $i = 1, 2, \dots$ — моменты окончания актов обслуживания требований, κ_i — число требований в очереди в момент τ_i , χ_i — состояние случайной среды на промежутке $[\tau_i, \tau_{i+1})$. Введем обозначения: $\beta_1 = \int_0^\infty t dB(t)$,

$\mu^{(k)} = \left. \frac{df^{(k)}}{dv} \right|_{v=1}$. В работах [1, 2] доказано, что стохастическая последовательность $\{(\kappa_i, \chi_i); i = 0, 1, \dots\}$ при заданном начальном распределении вектора (κ_0, χ_0) является однородной цепью Маркова. Данная цепь Маркова имеет единственное стационарное распределение тогда и только тогда, когда выполнено неравенство

$$\frac{a_{2,1}}{a_{1,2} + a_{2,1}} \cdot \frac{\lambda^{(1)} \mu^{(1)} \beta_1}{1 - p} + \frac{a_{1,2}}{a_{1,2} + a_{2,1}} \cdot \frac{\lambda^{(2)} \mu^{(2)} \beta_1}{1 - p} < 1. \quad (1)$$

По своей структуре выражение в левой части неравенства (1) похоже на «загрузку», так же как и по критической роли для существования стационарного режима. Будет ли это реальной загрузкой?

3. Исследование периода занятости

Под *периодом занятости* системы массового обслуживания понимают промежуток времени, начинающийся при поступлении в пустую систему первого требования и завершающийся, когда по окончании обслуживания система свободна от требований. Выберем дисциплину очереди LIFO — обратную порядку поступления требований. Тогда с каждым требованием можно связать свой период занятости, который начинается, когда требование начинает обслуживаться, и который заканчивается, когда завершается обслуживание всех требований, поступивших после начала обслуживания этого требования. Обозначим период занятости, связанный с j -м требованием, через \mathcal{Z}'_j , $j = 1, 2, \dots$.

Длительность i -го такта $\tau_i - \tau_{i-1}$, $i = 1, 2, \dots$, есть сумма двух слагаемых, $\mathfrak{v}_{1,i}$ и $\mathfrak{v}_{2,i}$. Величина $\mathfrak{v}_{2,i}$ всегда имеет функцию распределения $B(t)$ и не зависит от остальных. Величина $\mathfrak{v}_{1,i}$ при $\kappa_{i-1} > 0$ равна нулю, а при $\kappa_i = 0$ и $\chi_i = e^{(k)}$ имеет показательное распределение с параметром $\lambda^{(k)}$. Введем последовательность

$$\theta(1) = \inf\{i: i \geq 1, \kappa_i = 0\}, \quad \theta(j+1) = \inf\{i: i \geq \theta(j) + 1, \kappa_i = 0\}, \quad j = 1, 2, \dots$$

марковских моментов относительно последовательности $\{(\kappa_i, \chi_i); i = 0, 1, \dots\}$. Удобно считать, что $\inf \emptyset = \infty$ по определению. Моменты $\tau_i^{(3)} = \tau_{\theta(i)}$ при $\theta(i) < \infty$, $\tau_i^{(3)} = \infty$ при $\theta(i) = \infty$, $i = 1, 2, \dots$, суть моменты окончания периодов занятости. Длительность \mathcal{Z}_{i+1} периода занятости с порядковым номером $(i+1)$ в силу порядка обслуживания LIFO равна

$$\mathcal{Z}_{i+1} = \mathfrak{v}_{2,\theta(i)+1} + \mathfrak{v}_{2,\theta(i)+2} + \dots + \mathfrak{v}_{2,\theta(i+1)}, \quad i = 1, 2, \dots$$

Число слагаемых в этой сумме также случайно. Распределение величины \mathcal{Z}_{i+1} при условии $\{\kappa_{\theta(i)} = 0, \chi_{\theta(i)} = e^{(k)}\}$ совпадает с распределением величины \mathcal{Z}_1 при условии $\{\kappa_0 = 0, \chi_0 = e^{(k)}\}$. Предположим, что в начале очередь пуста, $\kappa_0 = 0$.

Не уменьшая общности, рассмотрим первый период занятости. Он начинается в момент поступления первой группы из $\eta_{1,1}$ требований.

Рассмотрим моменты окончания периодов занятости, связанных с поступившими в момент $\theta_{1,1}$ требованиями. Пусть

$$\begin{aligned}\theta(1, 0) &= 0, \quad \theta(1, 1) = \min\{i : i \geq 1, \kappa_i = \eta_{1,1} - 1\}, \\ \theta(1, j) &= \min\{i : i \geq \theta(1, j-1), \kappa_i = \kappa_{\theta(1, j-1)} - 1\}, \quad j = 2, 3, \dots, \eta_{1,1}.\end{aligned}$$

Случайная величина $\theta(1, j)$ задает номер такта, на котором заканчивается период занятости, связанный с j -м первичным требованием из первой группы. Обозначим

$$\hat{g}_{k,l}(s) = \mathbf{M}(e^{-s\beta'_1} I(\{\chi_{\theta(1,1)} = e^{(l)}\}) \mid \{\chi_{\theta(1,0)} = e^{(k)}\}), \quad k, l = 1, 2.$$

Введем матрицу

$$\hat{g}(s) = \begin{pmatrix} \hat{g}_{1,1}(s) & \hat{g}_{1,2}(s) \\ \hat{g}_{2,1}(s) & \hat{g}_{2,2}(s) \end{pmatrix}.$$

Элемент $(\hat{g}^j(s))_{k,l}$, расположенный в k -й строке и l -м столбце j -й степени матрицы $\hat{g}(s)$, представляет собой преобразование Лапласа

$$\mathbf{M}(e^{-s(\beta'_1 + \dots + \beta'_j)} I(\{\chi_{\theta(1,j)} = e^{(l)}\}) \mid \{\chi_{\theta(1,0)} = e^{(k)}\}).$$

Определим величины $P_k(b; t)$, $b = 0, 1, \dots$ равенством $\exp\{\lambda^{(k)} t (f^{(k)}(v) - 1)\} = \sum_{w=0}^{\infty} v^w P_k(w, t)$. Введем обозначения $\pi_k(b, s) = \int_0^{\infty} e^{-st} P_k(b; t) dB(t)$, E — единичная матрица 2×2 .

Теорема 1. Имеет место соотношение:

$$\hat{g}(s) = \sum_{b=0}^{\infty} \begin{pmatrix} \pi_1(b; s) & 0 \\ 0 & \pi_2(b; s) \end{pmatrix} \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \hat{g}^b(s) ((1-p)E + p\hat{g}(s)). \quad (2)$$

Период занятости системы полностью определяется преобразованиями Лапласа

$$\mathfrak{g}_{k,l}(s) = \mathbf{M}(e^{-s\beta_1} I(\{\chi_{\theta(1)} = e^{(l)}\}) \mid \{\chi_0 = e^{(k)}\}) = \sum_{w=1}^{\infty} f_w^{(k)} \cdot (\hat{g}^w(s))_{k,l}.$$

Для решения нелинейных уравнений вида (2) для преобразований Лапласа в статье [3] изложен метод последовательных приближений на основе теории обобщенных сжимающих отображений [4]. Рассмотрим нелинейный оператор

$$F(Z) = \sum_{b=0}^{\infty} \begin{pmatrix} \pi_1(b; s) & 0 \\ 0 & \pi_2(b; s) \end{pmatrix} \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} Z^b ((1-p)E + pZ)$$

на банаховом пространстве $M = \{Z: Z \in \mathbb{C}^{2 \times 2}\}$ всех комплекснозначных квадратных матриц размерности два с нормой $\|Z\| = \max\{|Z_{1,1}|+|Z_{1,2}|, |Z_{2,1}|+|Z_{2,2}|\}$. Очевидно, $\hat{g}(s) \in M$, $\Re s \geq 0$. Пусть $D = \{Z: \|Z\| \leq 1\}$ — замкнутый шар в M единичного радиуса с центром в точке

$$\Theta = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Далее, пусть $D^0 = \{Z: \|Z\| < 1\}$ — открытый шар единичного радиуса с центром в Θ , ∂D — граница множества D . Легко видеть, что значения матрицы $\hat{g}(s)$ при $\Re s \geq 0$ не могут выходить из множества D . Тогда нелинейный оператор $F(Z)$ при $\Re s \geq 0$ определен для всех $Z \in D$, непрерывно дифференцируем по Фреше в D^0 и является обобщенно-сжимающим в полуплоскости $\Re s > \sigma_1$ для некоторого $\sigma_1 > 0$. Решение уравнения $Z = F(Z)$ (т.е. уравнения (2)) может быть найдено методом последовательных приближений с начальным значением $Z = \Theta$.

Достаточное условие (1) существования стационарного распределения последовательности $\{(\kappa_i, \chi_i); i = 0, 1, \dots\}$ гарантирует, что период занятости будет принимать только конечные значения (с вероятностью единица). Тогда, решая уравнение (2) при $s = 0$, найдем $\hat{g}_{k,l}(0) = \mathbf{P}(\{\chi_{1,1} = e^{(l)}, \mathcal{Z}'_1 < \infty\} \mid \{\chi_0 = e^{(k)}\})$, т.е. условные вероятности для состояний случайной среды в моменты окончания периодов занятости, порожденных отдельными требованиями. Затем средняя длительность периода занятости \mathcal{Z}'_1 , начинающегося при состоянии среды $e^{(k)}$, может быть найдена дифференцированием: $\mathbf{M}(Z'_1 \mid \{\chi_0 = e^{(k)}\}) = -\hat{g}'_{k,1}(0) - \hat{g}'_{k,2}(0)$.

4. Определение загрузки системы

Определим вспомогательный процесс $\{K(t); t \geq 0\}$ (от греческого $\kappa\alpha\tau\varepsilon\iota\lambda\eta\mu\mu\varepsilon\nu\circ\zeta$ — занятый) следующим образом. Положим для $t \in [\tau_i, \tau_{i+1})$

$$K(t) = \begin{cases} (e^{(k)}, \Gamma^{(0)}) & \text{при } \mathfrak{v}_{1,i+1} > 0, t < \tau_i + \mathfrak{v}_{1,i+1}, \chi_i = e^{(k)}; \\ (e^{(k)}, \Gamma^{(1)}) & \text{при } t \geq \tau_i + \mathfrak{v}_{1,i+1}, \chi_i = e^{(k)}. \end{cases}$$

Здесь символ $\Gamma^{(0)}$ означает состояние простоя для обслуживающего устройства, символ $\Gamma^{(1)}$ означает, что прибор занят обслуживанием требования. Процесс $\{K(t); t \geq 0\}$ будет полумарковским.

Теорема 2. Стационарные вероятности $\pi_{k,j}$ состояний $(e^{(k)}, \Gamma^{(j)})$, $k = 1, 2$, $j = 0, 1$, для вложенной цепи скачков процесса $\{K(t); t \geq 0\}$ имеют вид

$$\pi_{1,j} = \mathfrak{g}_{2,1}(0)/(2(\mathfrak{g}_{1,2}(0) + \mathfrak{g}_{2,1}(0))), \quad \pi_{2,j} = \mathfrak{g}_{1,2}(0)/(2(\mathfrak{g}_{1,2}(0) + \mathfrak{g}_{2,1}(0))).$$

По теореме 10.2 из Главы 4 в [5], существуют предельные вероятности

$$\lim_{t \rightarrow \infty} \mathbf{P}(K(t) = (e^{(k)}, \Gamma^{(j)})) = \frac{\pi_{k,j} m_{k,j}}{\pi_{1,0} m_{1,0} + \pi_{2,0} m_{2,0} + \pi_{1,1} m_{1,1} + \pi_{2,1} m_{2,1}},$$

где $m_{k,j}$ — среднее время пребывания в состоянии $(e^{(k)}, \Gamma^{(j)})$; в нашем случае $m_{k,0} = (\lambda^{(k)})^{-1}$, $m_{k,1} = -\mathfrak{g}'_{k,1}(0) - \mathfrak{g}'_{k,2}(0)$. Поэтому вероятность застать прибор занятым в пределе равна

$$\begin{aligned} \rho = & -(\mathfrak{g}_{2,1}(0)(\mathfrak{g}'_{1,1}(0) + \mathfrak{g}'_{1,2}(0)) + \mathfrak{g}_{1,2}(0)(\mathfrak{g}'_{2,1}(0) + \mathfrak{g}'_{2,2}(0)))(\mathfrak{g}_{2,1}(0) \times \\ & \times ((\lambda^{(1)})^{-1} - \mathfrak{g}'_{1,1}(0) - \mathfrak{g}'_{1,2}(0)) + \mathfrak{g}_{1,2}(0)((\lambda^{(2)})^{-1} - \mathfrak{g}'_{2,1}(0) - \mathfrak{g}'_{2,2}(0))). \end{aligned} \quad (3)$$

По закону больших чисел, доля времени, в течение которого прибор занят обслуживанием, $(\mathfrak{v}_{2,1} + \dots + \mathfrak{v}_{2,n})(\mathfrak{v}_{1,i} + \mathfrak{v}_{2,i} + \dots + \mathfrak{v}_{1,n} + \mathfrak{v}_{2,n})^{-1}$, сходится при $n \rightarrow \infty$ почти наверное к величине ρ . Естественно называть величину ρ также *загрузкой* системы.

На основании приведенных формул выполнены численные эксперименты по вычислению загрузки для различных законов распределения длительности обслуживания. В этих экспериментах значение ρ не совпадало с левой частью неравенства (1).

5. Заключение

В работе рассмотрено формальное определение понятия загрузки системы обслуживания как доли времени, в течение которого обслуживающее устройство занято обслуживанием. Приведен строгий вывод расчетных формул для однолинейной системы с повторными требованиями и входным потоком с переменной структурой. Показано, что более простую форму имеет критерий стационарности, основанный не на загрузке.

ЛИТЕРАТУРА

1. Зорин А.В. О периоде занятости системы с дважды стохастическим входным потоком // Вестник Нижегородского университета им. Н. И. Лобачевского. Серия Математика. 2005. Вып. 3. С. 43–53.
2. Зорин А.В. Предельные свойства счетной цепи Маркова с квазитеплицевой переходной матрицей // Обозрение прикладной и промышленной математики. 2011. № 6. ISSN 0869-8325. С. 839–851.
3. Purdue P. Non-linear matrix integral equations of Volterra type in queueing theory // Journal of Applied Probability. 1973. V. 10. P. 644–651.
4. Kirk W.A. Mappings of generalized contractive type // Journal of mathematical analysis and applications. 1970. V. 32. P. 567–572.
5. Janssen J., Manca R. Applied Semi-Markov processes. Springer, 2006.

УДК: 004.942

Математическая модель двухпотоковой системы передачи данных

А.В. Рындин², И.А. Туренова¹, С.П. Моисеева¹, Е.А. Пакулова²

¹Национальный исследовательский Томский государственный университет, просп.
Ленина д.36, г. Томск, Россия

²Южный федеральный университет, ул. Большая Садовая ул., 105/42,
Ростов-на-Дону, Россия

Аннотация

В настоящей статье предлагается обобщенная модель Эрланга с ожиданием, особенность которой заключается в том, что для обработки и передачи данных предоставлены два гетерогенных ресурса конечного объема и разной интенсивности обслуживания. Для численного решения полученной из анализа диаграммы переходов процесса системы линейных уравнений равновесия для стационарных вероятностей предложен оригинальный рекуррентный скалярно-векторный алгоритм, позволяющий получить распределение вероятностей числа занятых приборов в каждом блоке и числа заявок, находящихся в очереди, а также рассчитать технические характеристики системы: загрузку каналов, среднее время пребывания в системе, вероятность моментального обслуживания, характеристики задержки в обслуживании и среднее число отказов в обслуживании.

Ключевые слова: гетерогенная система массового обслуживания, модель Эрланга, мультипотоковая система

1. Введение

Модель Эрланга широко используется при решении задач планирования объема инфраструктуры действующих и перспективных сетей связи. Основные положения модели разработаны более ста лет назад [1]. С момента ее появления произошли значительные изменения в разнообразии и распространении телекоммуникационных сервисов, в развитии сетевой инфраструктуры и технологий передачи информации, однако расчетные методики, полученные с использованием данной вероятностной конструкции, по-прежнему остаются в арсенале средств проектирования систем связи [2].

В настоящей статье предлагается обобщенная модель Эрланга с ожиданием, особенность которой заключается в том, что для обработки и передачи данных

представлены два гетерогенных ресурса конечного объема и разной интенсивности обслуживания. Под ресурсами понимаем в данном случае значения пропускной способности доступных каналов связи. Каналы связи, построенные на технологиях WiFi (первый канал) и 4G (второй канал), например, будут обладать гетерогенными характеристиками качества обслуживания (QoS). Таким образом, можно говорить об организации двухпотоковой системы передачи данных, так как информация приложения пользователя может быть передана по двум интерфейсам связи в зависимости от их загрузки. Реализация двупотоковой передачи данных возможна благодаря применению протокола мультипотоковой передачи данных, например, MPTCP [8]. Для измерения пропускной способности каналов связи, а также для определения скорости передачи данных вводится понятие единицы канального ресурса (ЕКР) каждого типа, используемое для обслуживания поступающих сообщений. Чаще всего ресурс, разделяемый между пользователями, – битовая скорость передачи информации. Тогда в качестве единицы ресурса берется наибольший общий делитель целочисленных аппроксимаций значений скорости канала связи. Общее число единиц каждого из ресурсов задают его объем, выраженный для удобства моделирования в целых числах.

Предполагается, что поступившее сообщение может использовать любую свободную ЕКР одного из каналов передачи информации, и все они идентичны с точки зрения возможностей и качества предоставления запрашиваемого сервиса. Дисциплина обслуживания определяется следующим образом. При поступлении сообщения сначала обращаются в первый канал, если он обладает достаточным количеством ЕКР, то сообщение попадает на обслуживание. В противном случае при недостаточном количестве свободного ресурса для обслуживания сообщение обращается во второй канал. Если второй канал обладает достаточным количеством свободного ресурса для его обслуживания, тогда сообщение попадает на обслуживание. В противоположном случае сообщение попадает в очередь. Дисциплина обслуживания в очереди FIFO (First In – First Out). Эта дисциплина называется "первый пришел – первый на обслуживание". Кроме того, сообщения, находящиеся в очереди, имеют срок жизни, после которого передавать их нет смысла.

Ставится задача определения характеристик качества обслуживания. А именно, загрузку каналов, среднее время пребывания в системе, вероятность моментального обслуживания, характеристики задержки в обслуживании и среднее число отказов в обслуживании.

2. Математическая модель

В качестве математической модели рассмотрим систему массового обслуживания (СМО) с двумя блоками обслуживания различной интенсивности обслужива-

ния на вход которой поступает пуссоновский поток заявок с интенсивностью λ . Предполагается, что канал включает в себя N_1 и N_2 единиц канального ресурса. Время обслуживания каждого сообщения является неотрицательной случайной величиной, имеющей экспоненциальное распределение с параметрами μ_1 и μ_2 соответственно. Если свободных ресурсов нет, то заявка поступает в очередь. Будем считать, что ограничений на очередь нет, но находящиеся в ней заявки могут покинуть ее. Время пребывания в очереди является случайной величиной, имеющей экспоненциальное распределение с параметром α .

Используя символику Д. Кендалла [3], запишем такую СМО в следующем виде: $M/(M_1, M_2)/(N_1, N_2)/\infty/FIFO$. Схематическое изображение системы представлено на рисунке 1.

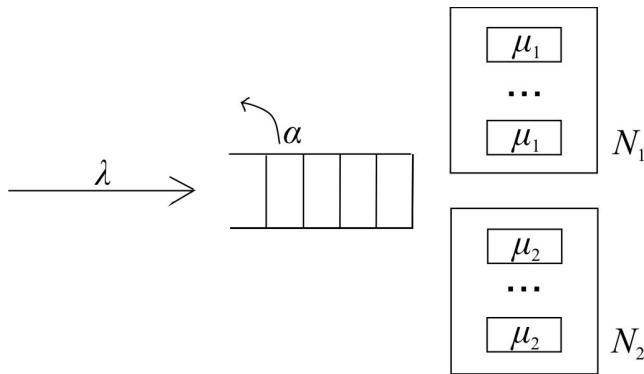


Рис. 1. Математическая модель

Следующий этап построения модели состоит в определении компонент случайного процесса, который будет использоваться для оценки показателей обслуживания заявок. Для решения поставленных задач оценки качества обслуживания необходимо знать среднее число занятых единиц канального ресурса в каждом блоке, вероятность постановки на ожидание (задержка передачи), средний размер очереди, долю потерянных заявок.

Обозначим $i_1(t)$ – число занятых единиц канального ресурса первого типа в момент времени t , $i_2(t)$ – число занятых единиц канального ресурса второго типа в момент времени t , $i_3(t)$ – число сообщений, находящихся в очереди в момент времени t . Определим $P(i, j, k, t) = P\{i_1(t) = i, i_2(t) = j, i_3(t) = k\}$ – вероятность того, что в первом блоке занято i линий, во втором j линий, в очереди k заявок, где $i = \overline{0, N_1}$, $j = \overline{0, N_2}$, $k = \overline{0, \infty}$.

Трехмерный случайный процесс $\{i_1(t), i_2(t), i_3(t)\}$ является марковским. Ставится задача нахождения стационарного распределения вероятностей $\Pi(i, j, k)$.

Графическая иллюстрация интенсивностей и направлений переходов из произвольного состояния для рассматриваемого процесса показана на рисунках 2, 3.

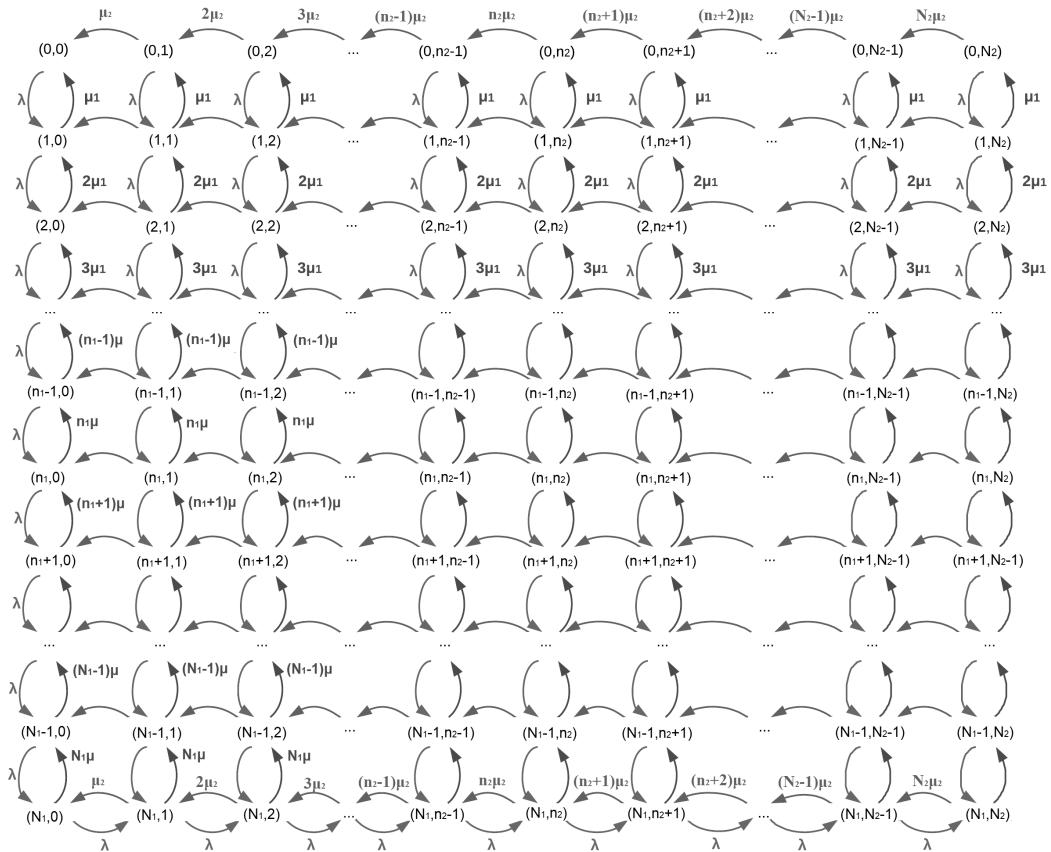


Рис. 2. Диаграмма переходов при $i_3(t) = 0$

Для численного решения полученной из анализа диаграммы переходов процесса системы линейных уравнений равновесия для стационарных вероятностей предложен оригинальный рекуррентный скалярно-векторный алгоритм, с помощью которого получены допредельные характеристики системы.

Выражения для вычисления показателей качества обслуживания заявок следуют из их физического смысла и определяются через отношение интенсивностей анализируемых событий или суммирование стационарных вероятностей модели, к ним относятся: среднее число заявок, ожидающих обслуживания; вероятность

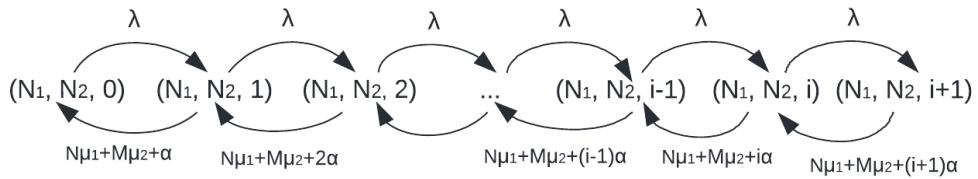


Рис. 3. Диаграмма переходов при $i_3(t) \neq 0$

попадания на ожидание; доля потерянных заявок; среднее время пребывания заявки на ожидании и в системе.

3. Численный анализ

Для вычисления показателей качества обслуживания заявок при моделировании были приняты следующие условия: время жизни пакетов принято $1/\alpha = 100$ мс, что соответствует трафику реального времени, интенсивность прибытия пакетов λ составляет 4,5 Мбит/сек, что соответствует передаче видео с разрешением 1080р (при применении H.264). Объем ресурсов сетей N_1, N_2 варьируются от 1 Мбит/сек до 4 Мбит/сек (см. табл. 1)

Характеристика	Объем ресурсов каналов связи, Мбит/сек		
	$N_1 = 4$	$N_1 = 2$	$N_1 = 1$
$N_2 = 4$	$N_2 = 4$	$N_2 = 4$	$N_2 = 2$
Среднее количество заявок в очереди	0.1467	0.5741	10.13
Интенсивность ухода из очереди	0.0147	0.05741	1.013
Вероятность потерь заявок	0.4%	1,4%	25%
Среднее время пребывания заявки в СМО, сек	0.1617	0.3102	2.866
Среднее время пребывания заявки в очереди, сек	0.0367	0.1435	2.533

Таблица 1. Характеристики СМО для системы двухпотоковой передачи данных

4. Заключение

В настоящей статье представлена математическая модель двухпотоковой системы передачи данных в виде гетерогенной модели Эрланга с ожиданием и нетерпеливыми заявками. С помощью разработанного рекуррентного алгоритма получено трехмерное распределение вероятностей (числа занятых приборов в каждом блоке и числа заявок, находящихся в очереди), и найдены технические характеристики системы, имеющие практическое значение для проектирования реальных информационных и телекоммуникационных систем. Вместе с тем следует отметить, что представленный алгоритм работает только для марковской модели, поэтому альтернативным подходом является применение метода асимптотического анализа, что позволяет найти асимптотическое распределение вероятностей числа заявок. В дальнейших работах также планируется обобщение результатов для модели с разноприоритетными входящими потоками.

5. Благодарности

Исследование было выполнено при поддержке гранта РФФИ «Принципы передачи многомодальной информации в роботизированных системах» (№19-37-90129).

Литература

1. Erlang A. K. The Theory of Probabilities and Telephone Conversations // Nyt Matematisk Tidsskrift. 1909. B. V. 20. P. 33-40.
2. Iversen V. B. Teletraffic Engineering and Network Planning. - Tehnial University of Denmark, May 2010. 370 p.
3. Kendall D. G. Some problems in the theory of queues // Journal of Royal Statistial Soiety. –1951. – Series B. – V. 13. –N. 2. – P. 151–173.
4. Vishnevsky V. M., Lyakhov A. I. IEEE 802.11 Wireless LAN: Saturation Throughput Analysis with Seizing Effect Consideration // Cluster Computing. 2002. V. 5. P. 133–144.
5. Neuts M. F. Structured Stochastic Matrices of M/G/1 Type and Their Applications. Marcel Dekker, New York, 1989.
6. Schriber T. J. Simulation using GPSS. John Wiley & Sons, 1974.
7. National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov>
8. Bonaventure O., Catholique de Louvain U., Paasch C. TCP Extensions for Multipath Operation with Multiple Addresses. – 2020.

65

УДК: 519.872.4

«Гауссовская аппроксимация для ресурсной гетерогенной СМО $(GI + 2M)^{(2,\nu)} / GI(2) / \infty$ »

Т.В. Бушкова¹ and С.П. Моисеева²

¹Национальный исследовательский Томский государственный университет, просп. Ленина д. 36, г. Томск, Россия
bushkova70@mail.ru, smoiseeva@mail.ru

Аннотация

В работе рассматривается ресурсная гетерогенная система массового обслуживания, состоящая из двух узлов, неограниченной ресурсной емкости. На вход поступают запросы на предоставление некоторого случайного объема ресурсов на некоторое случайное время. Потоки требований являются стационарными пуссоновскими различной интенсивности. Если для обслуживания заявки требуется задействовать ресурс обоих узлов, то предполагается, что моменты прихода таких заявок образуют рекуррентный поток с разделением на два разнотипных запроса. Отличительной особенностью рассматриваемых систем является то, что ресурс освобождается в запрашиваемом объеме. Для построения процесса изменения суммарно занимаемых ресурсов использован метод асимптотического анализа при условии эквивалентно растущей нагрузки. Построена гауссовская аппроксимация стационарного двумерного распределения вероятностей занятых ресурсов в узлах обслуживания.

Ключевые слова: ресурсные системы обслуживания, случайный объем требований

1. Введение

Интерес к ресурсным системам массового обслуживания (РСМО) объясняется актуальностью их применения для моделирования широкой области технических устройств и информационно-вычислительных систем. В отличие от классических СМО, где приборы и места ожидания играют роль ресурсов, необходимых для обслуживания заявок, в РСМО предполагается, что заявкам требуется случайный объем некоторых ресурсов. Рост популярности исследований таких систем обусловлен необходимостью создания эффективных инструментов оценки работы

сетей связи нового поколения, которые должны справляться с более сложной многопользовательской средой и использованием каналов на более высоких частотах [1]. К настоящему времени опубликовано большое количество работ, анализирующих РСМО [2].

Кроме того, при исследовании многолинейных СМО обычно предполагается, что серверы идентичны и поступающие требования занимают произвольный прибор для своего обслуживания. Менее изучены СМО с разнородными серверами, являющимися более интересным объектом для исследования [3]. При этом возможна ситуация, когда для входящего требования создается копия, передаваемая по другому каналу связи. В этом случае в качестве математической модели можно использовать многоресурсных СМО с параллельным обслуживанием.

Для исследования РСМО сегодня не существует универсальных способов, поэтому в работе применяются методы асимптотического анализа [4], проводящегося при условии эквивалентно растущего времени обслуживания [6] в совокупности с методом динамического просеивания, что позволяет получить приемлемые для практического использования асимптотические выражения искомых характеристик системы в случаях, когда их допредельное исследование невозможно.

2. Математическая модель

Рассмотрим РСМО с двумя узлами неограниченной емкости. На вход системы поступает три потока: два пуссоновских с параметрами λ_1 и λ_2 и рекуррентный поток заявок, длины интервалов между наступлениями событий в котором имеют функцию распределения вероятностей $A(z)$.

Предполагается, что в момент поступления в систему заявки входящих потоков формируют запросы на выделение некоторого ресурса, объем которого является неотрицательной случайной величиной $v_i \geq 0$, $i = 1, 2$, с функцией распределения вероятностей $G_i(x) = P\{v_i < x\}$, $i = 1, 2$, имеющей конечные первые и вторые моменты. Предполагается, что заявки простейших потоков формируют запросы на один ресурс соответствующего типа, а заявка рекуррентного потока – на ресурсы обоих типов. Каждый ресурс предоставляется на случайное время, с заданными функциями распределения вероятностей $B_k(x)$ ($k = 1, 2$), соответственно. Будем называть время, на которое предоставляется ресурс - временем обслуживания. Занятые ресурсы могут соответствовать ресурсам канала, захваченным сеансом передачи (например, радиочастоты, полоса пропускания и т.д.). После завершения обслуживания запроса ресурс освобождается в том же объеме, что и был занят. Ставится задача анализа общего количества занятых ресурсов в обоих каналах.

Обозначим $V_i(t)$ – суммарный объем ресурсов в системе, занятых заявками i -го типа ($i = 1, 2$) в момент времени t .

Для дальнейших исследований воспользуемся методом динамического просеивания [4]. Обозначим $S_i(t) = 1 - B_i(T - t)$ - динамические вероятности того, что заявка входящего потока, поступившая в i -ый блок системы в момент времени t , к некоторому произвольному фиксированному моменту времени $T > t$ не освободит ресурс, и будет находиться в системе. Соответственно, заявка закончит обслуживание и освободит занимаемый ресурс до момента времени T с вероятностью $1 - S_i(t) = B_i(T - t)$. Обозначим через $W_i(t)$ объемы просеянных ресурсов i -го типа, что будет соответствовать суммарному объему требований, не закончивших свое обслуживание к моменту времени T . Как показано в [7], законы распределения вероятностей значений случайного процесса $\mathbf{W}(t) = \{W_1(t), W_2(t)\}$ и исходного процесса $\mathbf{V}(t) = \{V_1(t), V_2(t)\}$ в момент времени $t = T$ совпадают:

$$P\{V_1(T) < x_1, V_2(T) < x_2\} = P\{W_1(T) < x_1, W_2(T) < x_2\} \quad \forall x_1, x_2. \quad (1)$$

Таким образом, исходная задача сводится к исследования двумерного процесса "просеянных объемов" $\{W_1(t, T), W_2(t, T)\}$. [5].

Поскольку на вход кроме пуссоновских поступает еще и рекуррентный поток, построенный процесс не будет являться марковским. Поэтому вводим дополнительную компоненту $z(t)$ - случайный процесс, описывающий длину интервала от момента времени t до момента наступления следующего события в рекуррентном потоке

Построенный трёхмерный процесс $\{z(t), W_1(t, T), W_2(t, T)\}$ является марковским, и для его распределения вероятностей $P(z, w_1, w_2, t) = P\{z(t) = z, W_1(t, T) < w_1, W_2(t, T) < w_2\}$ можно составить интегро-дифференциальное уравнение Колмогорова вида:

$$\begin{aligned} \frac{\partial P(z, w_1, w_2, t)}{\partial t} &= \frac{\partial P(z, w_1, w_2, t)}{\partial z} + \\ &[A(z)(1 - S_1(t) - S_2(t) + S_1(t)S_2(t)) - 1] \frac{\partial P(0, w_1, w_2, t)}{\partial z} + \\ &(-\lambda_1 S_1(t) - \lambda_2 S_2(t))P(z, w_1, w_2, t) + \lambda_1 S_1 \int_0^{w_1} P(z, w_1 - y_1, w_2, t) dG_1(y_1) + \\ &\lambda_2 S_2 \int_0^{w_2} P(z, w_1, w_2 - y_2, t) dG_2(y_2) + \\ &A(z) \left[S_1(t)(1 - S_2(t)) \int_0^{w_1} \frac{\partial P(0, w_1 - y_1, w_2, t)}{\partial z} dG_1(y_1) + \right. \end{aligned} \quad (2)$$

$$(1 - S_1(t))S_2(t) \int_0^{w_2} \frac{\partial P(0, w_1, w_2 - y_2, t)}{\partial z} dG_2(y_2) + \\ S_1(t)S_2(t) \left[\int_0^{w_1} \int_0^{w_2} \frac{\partial P(0, w_1 - y_1, w_2 - y_2, t)}{\partial z} dG_1(y_1) dG_2(y_2) \right]$$

Определим функции $h(z, u_1, u_2, t) = \int_0^{\infty} e^{ju_1 w_1} \int_0^{\infty} e^{ju_2 w_2} P(z, dw_1, dw_2, t)$, где $j = \sqrt{-1}$, и используя обозначения $G_k^*(u_k) = \int_0^{\infty} e^{ju_k y_k} dG_k(y_k)$, $k = 1, 2$, запишем уравнение:

$$\frac{\partial h(z, u_1, u_2, t)}{\partial t} = \frac{\partial h(z, u_1, u_2, t)}{\partial z} + h(z, u_1, u_2, t) [\lambda_1 S_1(t)(G_1^*(u_1) - 1) + \lambda_2 S_2(t)(G_2^*(u_2) - 1)] + \frac{\partial h(0, u_1, u_2, t)}{\partial z} \left\{ A(z) - 1 + A(z) \left[S_1(t)(G_1^*(u_1) - 1) + S_2(t)(G_2^*(u_2) - 1) + S_1(t)S_2(t)(G_1^*(u_1) - 1)(G_2^*(u_2) - 1) \right] \right\}. \quad (3)$$

$$h(z, u_1, u_2, t_0) = r(z). \quad (4)$$

Здесь $r(z)$ - стационарное распределение вероятностей процесса $z(t)$,

$$r(z) = \lambda \int_0^z (1 - A(x)) dx, \quad a = \int_0^{\infty} (1 - A(x)) dx, \quad \lambda = \frac{1}{a}. \quad (5)$$

Поскольку явный вид решения задачи (3) - (4) найти затруднительно, предлагается искать решение при асимптотическом условии эквивалентного роста времени обслуживания в блоках системы.

3. Асимптотический анализ при условии эквивалентно растущего времени

Обозначим среднее время обслуживания заявки в каждом блоке: $b_i = \int_0^{\infty} (1 - B_i(x)) dx$, $i = 1, 2$, тогда асимптотическое условие пропорционального роста времени обслуживания будет иметь вид $b_i \rightarrow \infty$, $i = 1, 2$, и $\lim_{b_i \rightarrow \infty} \frac{b_1}{b_2} = q = \text{const}$. Сформулируем без доказательства теорему об асимптотическом среднем рассматриваемого процесса.

Теорема 1. Асимптотическая характеристическая функция первого порядка $h^{(1)}(z, u_1, u_2, t) \approx h(z, u_1, u_2, t)$ для процесса $\{z(t), W_1(t), W_2(t)\}$ при условии эквивалентно растущего времени обслуживания имеет вид

$$h^{(1)}(z, u_1, u_2, t) = r(z) \exp \left\{ (\lambda + \lambda_1) j u_1 a_1 \int_{t_0}^t S_1(\xi) d\xi + (\lambda + \lambda_2) j u_2 a_2 \int_{t_0}^t S_2(\xi) d\xi \right\}, \quad (6)$$

где $a_i = \int_0^\infty y dG_i(y)$ - средний объем запроса на выделение ресурса в i -м блоке.

Полученная аппроксимация определяет математическое ожидание суммарных объемов занимаемых ресурсов каждого типа.

Более подробно остановимся на основных этапах построения аппроксимации второго порядка.

Функцию $h(z, u_1, u_2, t)$ определим как произведение вида:

$$h(z, u_1, u_2, t) = h^{(2)}(z, u_1, u_2, t) h^{(1)}(z, u_1, u_2, t), \quad (7)$$

Определив $b_1 = 1/\varepsilon^2$, $b_2 = 1/q\varepsilon^2$ и выполнив замены:

$$\varepsilon^2 t = \tau, S_i(t) = \bar{S}_i(\tau), u_i = \varepsilon x_i, h^{(2)}(z, u_1, u_2, t) = F_2(x_1, x_2, \tau, \varepsilon), \quad (8)$$

получаем дифференциальное уравнение вида:

$$\begin{aligned} & \varepsilon^2 \frac{\partial F_2(z, x_1, x_2, \tau, \varepsilon)}{\partial \tau} + F_2(z, x_1, x_2, \tau, \varepsilon) \left[(\lambda + \lambda_1) j \varepsilon x_1 a_1 + (\lambda + \lambda_2) j \varepsilon x_2 a_2 \bar{S}_2(\tau) \right] = \\ & = \frac{\partial F_2(z, x_1, x_2, \tau, \varepsilon)}{\partial z} + F_2(z, x_1, x_2, \tau, \varepsilon) \left[\lambda_1 \bar{S}_1(\tau) (G_1^*(u_1) - 1) + \lambda_2 \bar{S}_2(\tau) (G_2^*(u_2) - 1) \right] + \\ & + \frac{\partial F_2(0, x_1, x_2, \tau, \varepsilon)}{\partial z} (A(z) - 1) + \frac{\partial F_2(0, x_1, x_2, \tau, \varepsilon)}{\partial z} A(z) \left\{ \bar{S}_1(\tau) (G_1^*(u_1) - 1) + \right. \\ & \left. + \bar{S}_2(\tau) (G_2^*(u_2) - 1) + \bar{S}_1(\tau) \bar{S}_2(\tau) (G_1^*(u_1) - 1) (G_2^*(u_2) - 1) \right\} \end{aligned} \quad (9)$$

с начальным условием $F_2(z, x_1, x_2, \tau_0, \varepsilon) = r(z)$. решение которого будем искать в виде:

$$F_2(z, x_1, x_2, \tau, \varepsilon) = \Phi_2(x_1, x_2, \tau) \left\{ r(z) + j \varepsilon \left[\sum_{k=1}^2 (\lambda + \lambda_k) x_k a_k \bar{S}_k(\tau) \right] f_2(z) \right\} + O(\varepsilon^2).$$

Подставив его в (8), а также разделив на ε и устремив $\varepsilon \rightarrow 0$, имеем дифференциальное уравнение для нахождения функции $f_2(z)$:

$$\lambda(r(z) - A(z)) - f'_2(z) - f'_2(0)(A(z) - 1) = 0.$$

Собирая слагаемые при второй степени ε , получим дифференциальное уравнение для функции $\Phi_2(x_1, x_2, \tau)$, решение которого имеет вид:

$$\Phi_2(x_1, x_2, \tau) = \exp \left\{ j^2 \left(\frac{(x_1^2)}{2} (\lambda + \lambda_1) \alpha_1 \bar{S}_1(\tau) + 2(\lambda + \lambda_1) \kappa a_1^2 \bar{S}_1^2(\tau) + \right. \right. \\ \left. \left. + \frac{(x_2^2)}{2} (\lambda + \lambda_2) \alpha_2 \bar{S}_2^2(\tau) + 2(\lambda + \lambda_2) \kappa a_2^2 \bar{S}_2^2(\tau) + \frac{x_1 x_2}{2} \bar{S}_1(\tau) \bar{S}_2(\tau) a_1 a_2 \right) \right\}.$$

Подставляя это выражение в (9) и выполнив обратные замены, получаем:

$$h^{(2)}(z, u_1, u_2, t) = r(z) \exp \left\{ \frac{u_1^2}{2} \left((\lambda + \lambda_1) \alpha_1 S_1(t) + 2\kappa_1 (\lambda + \lambda_1) a_1^2 S_1^2(t) \right) + \right. \\ \left. + \frac{u_2^2}{2} \left((\lambda + \lambda_2) \alpha_2 S_2(t) + 2\kappa_1 (\lambda + \lambda_2) a_2^2 S_2^2(t) \right) + \right. \\ \left. + \frac{u_1 u_2}{2} S_1(t) S_2(t) a_1 a_2 \left[\lambda + 2\kappa \{(\lambda + \lambda_1) + (\lambda + \lambda_2) \} \right] \right\}, \quad (10)$$

где $\kappa = f'_2(0) = \lambda \int_0^\infty (r(x) - A(x)) dx$, $\alpha_k = \int_0^\infty y_k^2 dG_k(y_k)$, $k = 1, 2$,

Задав в (10) $z \rightarrow \infty$, $t = T$ и $t_0 \rightarrow -\infty$ получим выражение для характеристической функции совместного распределения занятых ресурсов в стационарном режиме:

$$h(u_1, u_2) = \exp \left\{ (\lambda + \lambda_1) j u_1 a_1 b_1 + (\lambda + \lambda_2) j u_2 a_2 b_2 + \right. \\ \left. \frac{u_1^2}{2} ((\lambda + \lambda_1) \alpha_1 b_1 + 2\kappa (\lambda + \lambda_1) a_1^2 \beta_1) + \frac{u_2^2}{2} ((\lambda + \lambda_2) \alpha_2 b_2 + 2\kappa (\lambda + \lambda_2) a_2^2 \beta_2) + \right. \\ \left. + \frac{u_1 u_2}{2} \beta_{12} a_1 a_2 [\lambda + 2\kappa \{(\lambda + \lambda_1) + (\lambda + \lambda_2)\}] \right\}, \quad (11)$$

где $\int_0^\infty [1 - B_k(x)]^2 dx \doteq \beta_k$, $\int_0^\infty [1 - B_1(x)] [1 - B_2(x)] dx \doteq \beta_{12}$,

Итак, распределение вероятностей двумерного процесса $\{V_1(t), V_2(t)\}$ является асимптотически гауссовским с вектором средних:

$$\mathbf{a} = [(\lambda + \lambda_1)a_1 b_1 \quad (\lambda + \lambda_2)a_2 b_2] \quad (12)$$

и матрицей ковариаций:

$$\mathbf{K} = \begin{bmatrix} (\lambda + \lambda_1)\alpha_1 b_1 + 2\kappa(\lambda + \lambda_1)a_1^2\beta_1 & \beta_{12}a_1 a_2[\lambda + 2\kappa\{(\lambda + \lambda_1) + (\lambda + \lambda_2)\}] \\ \beta_{12}a_1 a_2[\lambda + 2\kappa\{(\lambda + \lambda_1) + (\lambda + \lambda_2)\}] & (\lambda + \lambda_2)\alpha_2 b_2 + 2\kappa(\lambda + \lambda_2)a_2^2\beta_2 \end{bmatrix}. \quad (13)$$

4. Заключение

В данной статье построена гауссовская аппроксимация двумерного распределения вероятностей суммарных объемов занимаемых ресурсов в случае, когда на вход поступают два простейших потока заданной интенсивности и запросами на ресурсы разного типа и рекуррентный потока запросов на ресурсы обоих типов. Задача решена в предположении, что серверы имеют неограниченные ресурсы, а среднее время предоставления ресурсов, значительно больше среднего интервала между моментами поступления запросов.

Литература

- Горбунова А.В., Наумов В.А., Гайдамака Ю.В., Самуйлов К.Е. Ресурсные системы массового обслуживания с произвольным обслуживанием // Информатика и ее применения. 2019. Т. 13, вып. 1. С. 54–61.
- Наумов В. А. Самуйлов К. Е. Условия мультиплексивности стационарного распределения вероятностей марковских ресурсных систем массового обслуживания с потерями // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2019. № 46. С. 64–72.
- Efrosinin D.V., Farkhadov M.P., Stepanova N.V. Study of a Controllable Queueing System with Unreliable Heterogeneous Servers // Autom. Remote Control. 2018. V. 79. No. 2. P. 265–285.
- Назаров А.А., Моисеева С.П. Метод асимптотического анализа в теории массового обслуживания // Томск, Изд-во НТЛ 2006. С.109.
- Galileyskaya A., Lisovskaya E., Pagano M. On the Total Amount of the Occupied Resources in the Multi-resource QS with Renewal Arrival Process // CCIS. 2019. V. 1109. P. 257–269.
- Pankratova E.V. Heterogeneous system MMPP/GI(2)/∞ with random customers capacities /E.V. Pankratova, S.P. Moiseeva, M.P. Farhadov, A.N. Moiseev // Журн. СФУ. – Сер. Матем. и физ. – 2019. – 12:2 – p. 231–239
- Моисеев А. Н. Бесконечнолинейные системы и сети массового обслуживания / А. Н. Моисеев, А. А. Назаров. Томск: Изд-во НТЛ, 2015 –240 с.

УДК: 519.622.2

Исследование циклических систем с повторными вызовами в ключе построения сетей передачи данных

С.В. Пауль¹, К.С. Шульгина¹, О.Д. Лизюра¹, Д.В. Шашев¹

¹Национальный исследовательский Томский государственный Университет,
Томск, Российская Федерация

paulsv82@mail.ru, shulgina19991999@mail.ru, oliztsu@mail.ru, dshashev@mail.ru

Аннотация

В работе выполнено исследование математической модели циклической сети связи множественного доступа. Такая модель может использоваться в том числе для построения специализированных “летающих” сетей передачи данных FANET. В качестве математической модели такой сети рассматривается однолинейная система с повторными вызовами, на вход которой поступает N простейших потоков заявок, продолжительности обслуживания которых имеют экспоненциальную функцию распределения для заявок n -го потока, $n = 1 \dots N$. Применяя методы систем с прогулками прибора и асимптотического анализа в условии большой задержки заявок на орбите найдено асимптотическое распределение вероятностей значений числа заявок на выделенной орбите.

Ключевые слова: циклическая система, система с повторными вызовами, RQ-система, система с прогулками прибора

1. Введение

Специальные сети связи предназначены для обеспечения передачи данных между группой устройств, при этом в данной работе рассматриваются аспекты организации специальных сетей FANET (Flying Ad-Hoc Networks) [1] путем применения методов теории массового обслуживания в области построения моделей таких сетей. Такие сети предназначены для организации передачи данных в группе беспилотных летающих аппаратов (дронов).

Актуальной топологией для таких сетей связи может быть «звезда», центральный узел (ЦУ) которой выполняет функции управления группами дронов и является общим ресурсом сети. В качестве центрального узла может выступать

Исследование выполнено при поддержке Программы развития Томского государственного университета (Приоритет-2030).

центр управления группой устройств, центр управления полетами, диспетчерский пункт, взаимодействующий с группой дронов. Наличие сети позволяет каждому дрону передавать данные на центральный узел.

Проблема разделения общего ресурса связи сети может в подобных случаях решаться выбором протокола доступа абонентов сети к общему ресурсу.

Для эффективного разделения общего ресурса связи могут быть использованы циклические протоколы, либо протоколы множественного в том числе случайного доступа. При циклическом протоколе каждому дрону выделяется один временной интервал времени, в течение которого он полностью передает данные ЦУ. Временные окна следуют друг за другом, каждое временное окно закреплено за одним дроном. При использовании протокола случайного множественного доступа каждый дрон случайным образом выбирает окно для передачи информации, на которые разбит общий канал. Если несколько дронов выбирают одно и то же окно, то происходит столкновение (коллизия) данных. Повторная передача искаженных данных не производится. Неискаженная передача возможна при следующей передаче данных.

Проблема выбора класса протоколов эффективного доступа в данном исследовании решается методами математического моделирования и математически корректным исследованием предложенных моделей.

Адекватными математическими моделями протоколов случайного доступа являются RQ-системы (Retrial Queueing System) [2, 3, 4, 5], а для циклических протоколов системы поллинга [6, 7]. Те и другие являются математическими моделями систем массового обслуживания.

В данной работе предполагается исследовать вариант системы связи с группой дронов как системы массового обслуживания (СМО). Когда группа дронов находится на дежурстве, выполняет мониторинг местности или доставку груза, сеть связи группы находится в штатном циклическом режиме – каждый дрон передает собранную информацию в своем сегменте цикла центру управления (ЦУ). Будем рассматривать циклическую систему с повторными вызовами, в которой цикл является суммой интервалов доступа к общему ресурсу каждого из дронов. Особенностью предлагаемой модели является то, что продолжительности таких интервалов случайные (в частности, могут быть детерминированными) и независимыми не только между собой, но и не зависят от входящих потоков заявок и продолжительностей их обслуживания.

Ставится задача определения характеристик числа сообщений передаваемой информации в такой циклической системе массового обслуживания с повторными вызовами. Задача решается классическим методом «систем с прогулками прибора»[8, 9]. Указанный алгоритм назначения интервалов доступа независимых от входящего потока и времени обслуживания заявок решает основную

проблему систем поллинга. Для указанной стратегии эти времена независимы, поэтому многомерное распределение вероятностей факторизуется и, применяемый в работе метод «систем с прогулками» полностью решает поставленную задачу.

2. Математическая модель и постановка задачи

2.1. Математическая модель циклической системы с повторными вызовами. В процессе работы группы дронов, состоящей из N аппаратов, имеют место N пуассоновских потоков поступающих пакетов информации (заявок), передаваемых от дронов с интенсивностью λ_n для n -го дрона ($n = 1, \dots, N$) центру управления (рисунок 1).

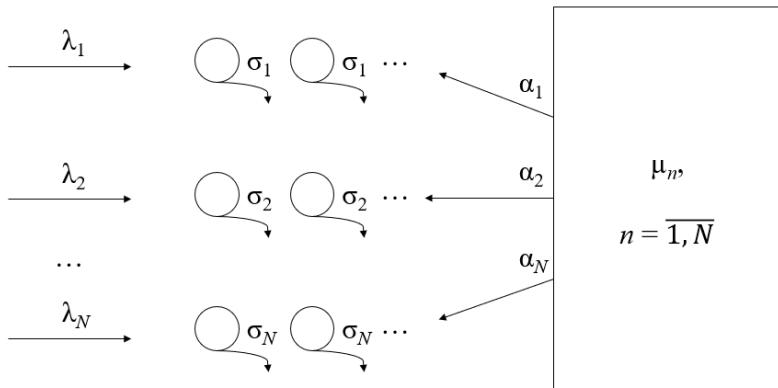


Рис. 1. Циклическая система с повторными вызовами

Заявки каждого потока формируют свою орбиту неограниченного объема. Пару поток и орбита будем называть n -ой RQ-системой. Прибор (ЦУ) посещает RQ-системы в циклическом порядке, начиная с первой и заканчивая N -ой, потом цикл повторяется. Это моделирует ситуацию, когда выделяется временной коридор для передачи информации от каждого дрона центру управления. Время нахождения прибора у n -ой RQ-системы имеет экспоненциальную функцию распределения с параметром α_n , $n = \overline{1, N}$. В течение этого времени прибор обслуживает заявки, которые поступают из входящего потока и с орбиты. Если поступившая заявка входящего потока обнаруживает прибор занятым или не подключенным, она мгновенно уходит на соответствующую орбиту, где осуществляет случайную задержку в течение экспоненциального времени с параметром σ_n , $n = \overline{1, N}$, после которой вновь обращается к прибору. Время обслуживания заявок имеет экспоненциальную функцию распределения с параметрами μ_n , $n = \overline{1, N}$.

Если заявок на орбите к моменту прихода прибора нет или он обслужил все заявки, которые находились на орбите, и из входящего потока больше не поступили новые заявки, прибор все равно остается подключенным к RQ-системе, пока не истечет время подключения. Методом исследования циклической системы является ее декомпозиция и исследование систем с прогулками прибора.

В результате исследования математической модели циклической системы будут получены аналитические выражения для параметров, определяющих улучшение качества функционирования сети связи группы дронов.

2.2. Метод прогулки прибора для циклической системы с повторными вызовами. Для исследования циклической системы с повторными вызовами перейдём к системе с прогулками прибора (рисунок 2).

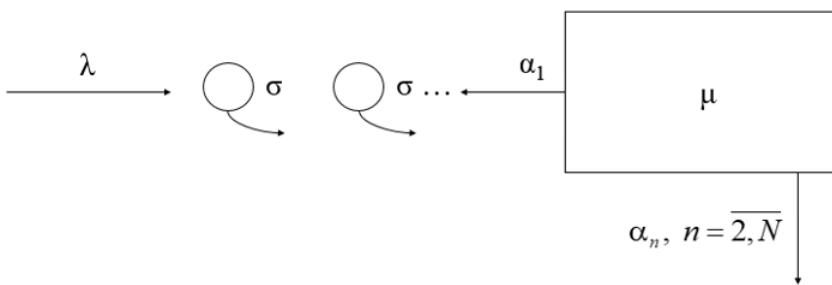


Рис. 2. Система с повторными вызовами с прогулками прибора

Рассмотрим первую систему с повторными вызовами с одним обслуживающим прибором и орбитой с неограниченным числом мест для ожидания. В систему поступает простейший поток заявок с интенсивностью λ . Система функционирует в циклическом режиме, цикл которой состоит из двух последовательных интервалов. В течение первого интервала прибор обслуживает заявки, которые поступают из входящего потока с экспоненциальной функцией распределения с параметром μ .

Если поступившая заявка из входящего потока обнаруживает прибор занятым, она мгновенно уходит на орбиту, где осуществляет случайную задержку в течение экспоненциального времени с параметром σ , после которой вновь обращается к прибору.

Продолжительность времени подключения прибора к потоку и орбите случайная и определяется экспоненциальной функцией распределения с параметром α_1 .

От момента окончания этого интервала прибор уходит на «прогулку», продолжительность которой складывается из $N - 1$ фаз. Каждая фаза имеет экспоненциальное распределение с параметрами $\alpha_n, n = \overline{2, N}$. Во время прогулки,

пришедшие в систему, заявки накапливаются на орбите и ждут, когда прибор вернется на обслуживание.

Обозначим процесс $k(t)$ – состояние прибора в момент времени t . Этот процесс может принимать следующие значения: 0 – прибор свободен, 1 – прибор занят обслуживанием заявки, n – прибор на n -ой фазе прогулки, $n = \overline{2, N}$.

Также введем случайный процесс $i(t)$ – число заявок на орбите в момент времени t .

2.3. Система дифференциальных уравнений Колмогорова. Ставится задача нахождения стационарного распределения числа заявок на орбите. Для этого рассмотрим двумерный марковский процесс $\{k(t), i(t)\}$. Для распределения вероятностей $P\{k(t) = k, i(t) = i\} = P_k(i, t)$, составим систему Колмогорова. Введем частичные характеристические функции, обозначив $j = \sqrt{-1}$: $H_k(u, t) = \sum_{i=0}^{\infty} e^{ju_i} P_k(i, t)$, $k = \overline{0, N}$. Сделаем замены, получим систему для функций $H_k(u, t)$, которую будем решать методом асимптотически-диффузационного анализа в предельном условии ($\sigma \rightarrow 0$).

Теорема 1. Стационарная плотность распределения нормированного и центрированного числа заявок на орбите имеет вид:

$$s(z) = \frac{C}{b(z)} \exp \left\{ \frac{2}{\sigma} \int_0^z \frac{a(x)}{b(x)} dx \right\}, \quad (1)$$

где C – нормирующая константа, $a(x)$ – определяется равенством:

$$a(x) = \lambda - \left(x - \frac{\alpha_1(\lambda + x)}{\mu + \alpha_1} + \lambda \right) r_0(x),$$

здесь $x = x(\tau)$ является решением уравнения:

$$x'(\tau) = -x(\tau)r_0 + \alpha_1 r_1 + \lambda \sum_{n=1}^N r_n,$$

вероятности $r = r_k(x)$, $k = \overline{0, N}$:

$$r_0(x) = \left[\frac{\mu + \alpha_1 + \lambda + x}{\mu + \alpha_1} + \sum_{n=2}^N \frac{\alpha_1(\mu + \alpha_1 + \lambda + x)}{\alpha_n(\mu + \alpha_1)} \right]^{-1}, \quad r_1(x) = \frac{\lambda + x}{\mu + \alpha_1} r_0(x),$$

$$r_2(x) = \frac{\alpha_1(\mu + \alpha_1 + \lambda + x)}{\alpha_2(\mu + \alpha_1)} r_0(x), \quad r_n(x) = \frac{\alpha_1(\mu + \alpha_1 + \lambda + x)}{\alpha_n(\mu + \alpha_1)} r_0(x), \quad n = \overline{3, N},$$

$b(x)$ - определяется равенством:

$$b(x) = a(x) + 2 \left(-xg_0(x) + \alpha_1 g_1(x) + \lambda \sum_{k=1}^N g_k(x) + xr_0(x) \right),$$

функции $g_k(x), k = \overline{0, N}$ определяются неоднородной системой:

$$\begin{aligned} -(\lambda + \alpha_1 + x)g_0 + \mu g_1 + \alpha_N g_N &= a(x)r_0, \\ -(\mu + \alpha_1)g_1 + (\lambda + x)g_0 &= a(x)r_1 - \lambda r_1 + xr_0, \\ -\alpha_2 g_2 + \alpha_1 g_1 + \alpha_1 g_0 &= a(x)r_2 - \lambda r_2 - \alpha_1 r_1, \\ -\alpha_n g_n + \alpha_{n-1} g_{n-1} &= a(x)r_n - \lambda r_n, \quad n = \overline{3, N}, \\ \sum_{k=0}^N g_k &= 0. \end{aligned}$$

3. Область применимости асимптотически-диффузационного анализа

Данный пункт посвящен определению границ применимости полученных аналитически-диффузационных результатов. Сравнивая асимптотические результаты с допредельным распределением, полученным ранее в работе [10], мы можем определить при каких значениях параметров асимптотически-диффузационное распределение вероятностей близко к допредельному.

Определив неотрицательную функцию $G(i)$ дискретного аргумента i в виде

$$G(i) = \frac{C}{b(\sigma i)} \exp \left\{ \frac{2}{\sigma} \int_0^{\sigma i} \frac{a(x)}{b(x)} dx \right\},$$

Построим аппроксимацию $P_{dif}(i)$ распределения вероятностей $P(i) = P\{i(t) = i\}$ числа i заявок на орбите для RQ-систем с использованием формулы (1)

$$P_{dif}(i) = \frac{G(i)}{\sum_{i=0}^{\infty} G(i)}.$$

В таблице 1 приведены значения этих расстояний для различных параметров σ .

	$\sigma = 0,1$	$\sigma = 0,07$	$\sigma = 0,05$	$\sigma = 0,03$	$\sigma = 0,01$
Δ	0,064	0,052	0,047	0,043	0,042

Таблица 1. Расстояние Колмогорова Δ

Анализируя данные таблицы 1, можно сказать, что точность аппроксимации растет с уменьшением параметра σ . Приведенные аппроксимации применимы для расстояния Колмогорова не превышающем значения 0,05. Полужирным в таблице 1 выделены те значения, при которых будем считать точность аппроксимаций удовлетворительными.

4. Заключение

В данной работе представлено исследование циклической системы с повторными вызовами относительно организации специальной сети передачи данных FANET. Анализ представленной модели выполнен методом асимптотически-диффузационного анализа. По полученным результатам была построена аппроксимация распределения вероятностей числа заявок на орбите. В главе численного анализа показана точность построенной аппроксимации асимптотически-диффузационного анализа.

Литература

1. Khan, M.F.; Yau, K.-L.A.; Noor, R.M.; Imran, M.A. Routing Schemes in FANETs: A Survey. Sensors 2020, 20, 38.
2. J. R. Artalejo, Accessible bibliography on retrial queues: progress in 2000–2009, Mathematical and computer modelling 51 (9-10) (2010) 1071–1081.
3. J. R. Artalejo, A classified bibliography of research on retrial queues: progress in 1990–1999, Top 7 (2) (1999) 187–211.
4. J. R. Artalejo, A. G’omez-Corral, Retrial queueing systems, Vol. 30, Springer, 1999.
5. J. R. Artalejo, Algorithmic Methods in Retrial Queues, Vol. 141, Springer, 2006.
6. В. М. Вишневский, О. В. Семенова, Системы поллинга: теория и применение в широкополосных беспроводных сетях. М.: Техносфера, 2007. 312 с., Информационные технологии и вычислительные системы (1) (2008) 98–99.
7. В. М. Вишневский, О. В. Семенова, Математические методы исследования систем поллинга, Автоматика и телемеханика (2) (2006) 3–56.
8. A. Nazarov, S. Paul, A cyclic queueing system with priority customers and strategy of service, Communications in Computer and Information Science 678 (2016) 182–193.

9. A. Nazarov, S. Paul, A number of customers in the system with server vacations, Communications in Computer and Information Science 601 (2015) 334–343.
10. Назаров А. А. Исследование циклической системы с повторными вызовами / А. А. Назаров, С. В. Пауль, П. Н. Ключникова // Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2020) [Электронный ресурс] : материалы XXIII Международной научной конференции (14-18 сентября 2020 г., Москва, Россия). М., 2020. С. 540-547.

UDC: 519.872

Analysis of Tandem Retrial Queue with Common Orbit and MMPP Incoming Flow

S.V. Paul¹, A.A. Nazarov¹, T. Phung-Duc², M.A. Morozova¹

¹National Research Tomsk State University, 36 Lenina ave., 634050, Tomsk, Russia

²University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

paulsv82@mail.ru, nazarov.tsu@gmail.com, tuan@sk.tsukuba.ac.jp,
morozova_mariya_a@mail.ru

Abstract

In this paper, we consider a tandem queueing system with one orbit, MMPP incoming flow and two sequentially connected servers by a method of asymptotic diffusion analysis involving the existence of a steady-state regime in condition of unlimited increase of the average delay time of calls in the orbit. We obtain the condition for the existence of the steady-state regime and the limiting probability distribution of the number of calls there. Then we evaluate the applicability of the asymptotic results by simulation.

Keywords: tandem system, retrial queue system, MMPP incoming flow, asymptotic diffusion analysis

1. Introduction

Retrial phenomenon naturally arises in various systems such as communication systems and service systems. For example, in call centers, if an operator is not available upon the arrival of a customer, the customer may hear some message such that “the system is currently congested, please wait or make a phone call later”. In computer systems, if a request is not processed in a period of time, some automatic program tries to repeat the request in some fixed intervals. From a queueing model point of view, in these situations customers who cannot receive service immediately upon arrival due to the unavailability of the servers join the orbit and retry to enter the server after some random time. During the retrial interval, customers stay in a virtual waiting room called the orbit.

Retrial queues characterized by the phenomenon above have been extensively studied in the literature [1]. The analysis of these models is more difficult than that of their counterpart models with a buffer in front of the server(s) because the underlying Markovian chain of retrial queues is spatially inhomogeneous due to the

total retrial rate of customers that depends on the number of customers in the orbit. As a reason, analytical results for retrial queues are available in only a few special cases with one or two servers [2].

The analysis of queueing network with retrials is even more challenging because these models do not possess a product form solution. There are several related works on tandem queue with two connected servers in which only blocked customers at the first server join orbit while those who find the second server busy upon service completion at the first server are lost. For such a model, in the cases with exponential service time distributions in both servers, explicit solution is derived [3]. For matrix-analytic solutions, some more general solutions are available [4].

However, for the models with retrials from any server, analytical solutions have not been obtained yet for even the simplest model with Poisson input. In our recent study [5], we have obtained a scaling limit for the exponential distribution setting at both servers. In this paper, we extend our previous work to present the solution for such a model with MMPP input. It should be noted that explicit solution for the joint stationary distribution of the number of customers in orbit and the state of the servers is challenging if not impossible.

Here we focus on a special regime when the retrial rate is extremely small. In this regime, the number of customers in orbit explodes. However, after a proper scaling, we prove that two scaled versions of the number of customers in the orbit converge to a deterministic process and a diffusion process respectively. The later result is then used to build an approximation to the distribution of the number of customers in orbit.

2. Description of the Mathematical Model and Problem Statement

We consider a retrial queueing system with MMPP arrivals and two sequentially connected servers (Figure 1). Upon the arrival of a call, if the first server is free, the call occupies it. The call is served for a random time exponentially distributed with parameter $/mu_1$ and then tries to go to the second server. If the second server is free, the call moves to it for a random time exponentially distributed with parameter $/mu_1$ and after that departs from the system. When a call arrives, if the first server is busy, the call instantly goes to the orbit, stays there during random time exponentially distributed with parameter σ and then retries to occupy the first server again. The behavior of a call from the orbit is the same as that of a new one.

Let us denote: process $k(t)$ – Markovian chain which manages the MMPP-flow, defined by infinitesimal generator – matrix \mathbf{Q} with elements q_{vk} , $v, k = \overline{1, K}$; the arrival rate is given by $\lambda_k \geq 0$; process $n_1(t)$ – the state of the first server at time t : 0, if the server is free; 1, if the server is busy; process $n_2(t)$ – the state of the second

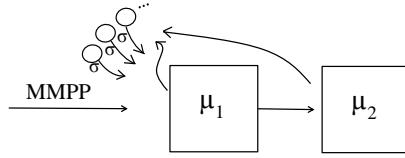


Fig. 1. Tandem Retrial Queue with common orbit and incoming MMPP-flow

server at time t : 0, if the server is free; 1, if the server is busy; process $i(t)$ – number of calls in the orbit at the time t .

The goal of the study is to obtain a probability distribution number of calls in the orbit $i(t)$ in the considered system.

3. The System of Differential Kolmogorov Equations

We define probabilities

$$P_{n_1 n_2}(k, i, t) = P\{n_1(t) = n_1, n_2(t) = n_2, k(t) = k, i(t) = i\},$$

$$n_1 = 0, 1; n_2 = 0, 1. \quad (1)$$

The four-dimensional process $\{n_1(t), n_2(t), k(t), i(t)\}$ is a Markovian chain. For probability distribution (1), we can write the system of differential Kolmogorov equations:

$$\begin{aligned} \frac{\partial P_{00}(k, i, t)}{\partial t} &= -(\lambda_k + i\sigma)P_{00}(k, i, t) + \mu_2 P_{01}(k, i, t) + \sum_v P_{00}(v, i, t)q_{vk}, \\ \frac{\partial P_{10}(k, i, t)}{\partial t} &= \lambda_k P_{00}(k, i, t) + (i+1)\sigma P_{00}(k, i+1, t) - (\lambda_k + \mu_1)P_{10}(k, i, t) + \\ &\quad + \lambda_k P_{10}(k, i-1, t) + \mu_2 P_{11}(k, i, t) + \sum_v P_{10}(v, i, t)q_{vk}, \\ \frac{\partial P_{01}(k, i, t)}{\partial t} &= \mu_1 P_{10}(k, i, t) - (\lambda_k + \mu_2 + i\sigma)P_{01}(k, i, t) + \\ &\quad + \mu_1 P_{11}(k, i-1, t) + \sum_v P_{01}(v, i, t)q_{vk}, \\ \frac{\partial P_{11}(k, i, t)}{\partial t} &= -(\lambda_k + \mu_1 + \mu_2 + i\sigma)P_{11}(k, i, t) + \\ &\quad + \lambda_k P_{11}(k, i, t) + \sum_v P_{11}(v, i, t)q_{vk}, \end{aligned} \quad (2)$$

We introduce partial characteristic functions, denoting $j = \sqrt{-1}$

$$H_{n_1 n_2}(k, u, t) = \sum_{i=0}^{\infty} e^{jui} P_{n_1 n_2}(k, i, t). \quad (3)$$

So, we have

$$\begin{aligned} \frac{\partial H_{00}(k, u, t)}{\partial t} &= -\lambda_k H_{00}(k, u, t) + \mu_2 H_{01}(k, u, t) \\ &\quad + \sum_v H_{00}(v, u, t) q_{vk} + j\sigma \frac{\partial H_{00}(k, u, t)}{\partial u}, \\ \frac{\partial H_{10}(k, u, t)}{\partial t} &= (\lambda_k (e^{ju} - 1) - \mu_1) H_{10}(k, u, t) + \lambda_k H_{00}(k, u, t) + \\ &\quad + \mu_2 H_{11}(k, u, t) + \sum_v H_{01}(v, u, t) q_{vk} - j\sigma e^{-ju} \frac{\partial H_{00}(k, u, t)}{\partial u}, \\ \frac{\partial H_{01}(k, u, t)}{\partial t} &= \mu_1 H_{10}(k, u, t) - (\lambda_k + \mu_2) H_{01}(k, u, t) + \mu_1 e^{ju} H_{11}(k, u, t) + \\ &\quad + \sum_v H_{10}(v, u, t) q_{vk} + j\sigma \frac{\partial H_{01}(k, u, t)}{\partial u}, \\ \frac{\partial H_{11}(k, u, t)}{\partial t} &= (\lambda_k (e^{ju} - 1) - \mu_1 - \mu_2) H_{11}(k, u, t) + \\ &\quad + \sum_v H_{11}(v, u, t) q_{vk} + j\sigma e^{-ju} \frac{\partial H_{01}(k, u, t)}{\partial u}. \end{aligned} \quad (4)$$

Let us denote row vectors to remain the compactness of further computation

$$\begin{aligned} \mathbf{H}_{n_1 n_2}(u, t) &= \{H_{n_1 n_2}(1, u, t), H_{n_1 n_2}(2, u, t), \dots, H_{n_1 n_2}(K, u, t)\} \\ \mathbf{H}(u, t) &= \{\mathbf{H}_{00}(u, t), \mathbf{H}_{10}(u, t), \mathbf{H}_{01}(u, t), \mathbf{H}_{11}(u, t)\} \end{aligned} \quad (5)$$

Furthermore, we define the following block structured matrices

$$\mathbf{A} = \begin{bmatrix} \mathbf{Q} - \Lambda & \Lambda & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{Q} - (\Lambda + \mu_1 \mathbf{I}) & \mu_1 \mathbf{I} & \mathbf{O} \\ \mu_2 \mathbf{I} & \mathbf{O} & \mathbf{Q} - (\Lambda + \mu_2 \mathbf{I}) & \Lambda \\ \mathbf{O} & \mu_2 \mathbf{I} & \mathbf{O} & \mathbf{Q} - (\Lambda + \mu_1 \mathbf{I} + \mu_2 \mathbf{I}) \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \Lambda & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mu_2 \mathbf{I} & \Lambda \end{bmatrix}, \mathbf{I}_0 = \begin{bmatrix} \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} \end{bmatrix}, \mathbf{I}_1 = \begin{bmatrix} \mathbf{O} & \mathbf{I} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{I} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} \end{bmatrix}. \quad (6)$$

There are all blocks with the dimension $K \times K$, \mathbf{O} is a zero block with the dimension $K \times K$, \mathbf{I} is identity matrix. Using these matrices and multiplying all equations of system 4 by vector of units \mathbf{e} with dimension $4K$, combining matrix equation and scalar equation we have the system

$$\begin{aligned} \frac{\partial \mathbf{H}(u, t)}{\partial t} &= \mathbf{H}(u, t)\{\mathbf{A} + e^{ju}\mathbf{B}\} + j\sigma \frac{\partial \mathbf{H}(u, t)}{\partial u}\{\mathbf{I}_0 - e^{-ju}\mathbf{I}_1\}, \\ \frac{\partial \mathbf{H}(u, t)}{\partial t}\mathbf{e} &= (e^{ju} - 1) \left\{ \mathbf{H}(u, t)\mathbf{B} + j\sigma e^{-ju} \frac{\partial \mathbf{H}(u, t)}{\partial u} \mathbf{I}_1 \right\} \mathbf{e}. \end{aligned} \quad (7)$$

This system of equations is the basis in further research. We solved it by a method of asymptotic diffusion analysis under the asymptotic condition $\sigma \rightarrow 0$ for similar tandem queueing system with Poisson arrival process in [5]. We obtained parameters of the diffusion process. Probability density distribution of this process has enabled us to construct an approximation for probability distribution number of calls in the orbit in the considered RQ-system.

4. Existence of Steady-State Regime

The inequality $\lim_{x \rightarrow \infty} a(x) < 0$ [[5]] is the necessary equation for the existence of the steady-state regime. Let us prove the following statement. We will use the following statement.

Theorem 1. A necessary condition for existence of steady-state regime in RQ-system under consideration is inequality

$$\mathbf{r}_1 \Lambda \mathbf{e}_1 < \frac{\mu_1 \mu_2}{\mu_1 + \mu_2}. \quad (8)$$

Vector \mathbf{r}_1 – is the vector of steady-state distribution of the control process $k(t)$ for which $\mathbf{r}_1 \mathbf{Q} = 0$, $\mathbf{r}_1 \mathbf{e}_1 = 1$, Λ – is the diagonal matrix with elements λ_k , $k = \overline{1, K}$ and vector \mathbf{e}_1 is an vector of units with dimension K .

This equality defines a convergence of characteristic functions $\sigma i(\tau/\sigma)$ normalized random process to deterministic function $x(\tau)$ under the condition $\sigma \rightarrow 0$.

5. Approximations Accuracy

We have constructed an approximation for discrete probability distribution $P(i)$ in [5]. We have written a non-negative function $G(i)$ of the discrete argument i in

the form

$$G(i) = \frac{C}{b(\sigma i)} \exp \left\{ \frac{2}{\sigma} \int_0^{\sigma i} \frac{a(x)}{b(x)} dx \right\}, \quad (9)$$

where function $b(x)$ is a diffusion coefficient of a diffusion process which has the function $a(x)$ as a coefficient of drift.

Using the normalization condition, we can write

$$P_1(i) = \frac{G(i)}{\sum_{i=0}^{\infty} G(i)}. \quad (10)$$

This probability distribution $P(i)$ we will use as an approximation for the probability distribution $P(i) = P\{i(t) = i\}$ that the number of calls in the orbit.

Approximations accuracy will be defined and compare by using Kolmogorov range

$$\Delta = \max_{k \geq 0} \left| \sum_{i=0}^k (P(i) - P(i)) \right|, \quad (11)$$

where $P(i)$ is an empirical probability distribution of the number i of calls in the orbit obtained by the simulation.

Let's denote matrices

$$\mathbf{Q} = \begin{bmatrix} -1 & 0.3 & 0.7 \\ 0.1 & -0.6 & 0.5 \\ 0.4 & 0.3 & -0.7 \end{bmatrix}, \mathbf{\Lambda}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Load of the system by ρ ($0 < \rho < 1$) and intensity values of the incoming flow – elements of matrix Λ (under the condition for the existence of steady-state regime (8)) set by en equation

$$\mathbf{\Lambda} = \rho \frac{\mathbf{\Lambda}_1}{\mathbf{r}_1 \mathbf{\Lambda}_1 \mathbf{e}_1} \frac{\mu_1 \mu_2}{\mu_1 + \mu_2}. \quad (12)$$

We consider $\mu_1 = 1$, $\mu_2 = 2$ and $\rho = 0.5$ for different parameters σ .

$\sigma = 2$	$\sigma = 1.5$	$\sigma = 1$	$\sigma = 0.5$	$\sigma = 0.1$	$\sigma = 0.05$	$\sigma = 0.01$
0.057	0.049	0.033	0.011	0.033	0.019	0.007

Table 1. Kolmogorov range.

It can be seen from the table that the accuracy of the approximations increases with decreasing parameter σ . The Gaussian approximation is applicable for values of $\sigma < 1.5$, where the relative error, in the form of the Kolmogorov distance, does not exceed 0.05.

6. Conclusion

In this paper, we consider the tandem retrial queueing system with incoming MMPP-flow. For this system, we define the condition for the existence of the steady-state regime. Using the method of asymptotic diffusion analysis under the asymptotic condition of the long delay in the orbit, we obtain parameters of the diffusion process. Probability density distribution of this process has enabled us to construct an approximation for probability distribution number of calls in the orbit in the considered retrial queueing system. Comparing with the results of simulation, it is shown that the accuracy of the approximations increases with decreasing parameters σ .

REFERENCES

1. G. I. Falin, J. Templeton, *Retrial Queues*, 1st Edition, Chapman & Hall, London, 1997.
2. T. Phung-Duc, H. Masuyama, S. Kasahara, Y. Takahashi, State-dependent $M/M/c/c + r$ retrial queues with bernoulli abandonment, *Journal of Industrial and Management Optimization* 6 (3) (2010) 517–540.
3. T. Phung-Duc, An explicit solution for a tandem queue with retrials and losses, *Operational Research*, 12 (2) (2012) 189–207.
4. C. S. Kim, V. Klimenok, O. Taramin, A tandem retrial queueing system with two markovian flows and reservation of channels, *Computers & operations research*, 37 (7) (2010) 1238–1246.
5. A. Nazarov, S. Paul, T. Phung-Duc, M. Morozova, Analysis of tandem retrial queue with common orbit and poisson arrival process, in: *Performance Engineering and Stochastic Modeling*, Springer, 2021, pp. 441–456.

УДК: 519.872

Изучение процесса адаптивного управления конфликтными потоками Кокса-Льюиса путем имитационного моделирования

Е.В. Кудрявцев¹ and М.А. Федоткин¹

¹ННГУ им. Н.И. Лобачевского, Нижний Новгород, Россия

evgkudryavcev@gmail.com, fma5@rambler.ru

Аннотация

В работе рассматривается система адаптивного управления конфликтными потоками разнотипных требований. Математической моделью является векторная марковская последовательность. Для данной последовательности получены условия существования предельного распределения. Также проведено численное исследование системы методом имитационного моделирования. Приведен алгоритм оптимизации имитационной модели.

Ключевые слова: конфликтные потоки, адаптивное управление, стационарный режим, имитационное моделирование

1. Введение

Исследуется система управления конфликтными неординарными пуассоновскими потоками [1]. В книге Д. Кокса и П. Льюиса «Статистический анализ последовательностей событий» [2] уделяется большое внимание потокам случайных событий (точечным процессам), в которых интервалы между наступлениями событий зависят от времени и имеют разное распределение. В частности, приводится большое число таблиц с реальными данными, обладающими такими свойствами. В работах [3, 4] было показано, что статистические данные из нескольких таблиц, приведенных Д. Коксом и П. Льюисом, могут быть аппроксимированы неординарными пуассоновскими потоками (потоками с пакетным поступлением). Потоки такого рода задаются интенсивностью λ потока вызывающих моментов

и распределением количества χ требований в вызывающий момент

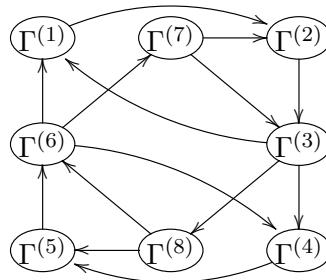
$$\begin{aligned}\mathbf{P}(\chi = 1) &= p = \frac{1}{1 + \alpha + \alpha\beta(1 - \gamma)^{-1}}, \\ \mathbf{P}(\chi = 2) &= \frac{\alpha}{1 + \alpha + \alpha\beta(1 - \gamma)^{-1}}, \\ \mathbf{P}(\chi = k) &= \frac{\alpha\beta\gamma^{k-3}}{1 + \alpha + \alpha\beta(1 - \gamma)^{-1}}, \quad k \geq 3,\end{aligned}\tag{1}$$

где α , β и γ — параметры потока.

2. Математическая модель системы

Два независимых конфликтных неординарных пуассоновских потока Π_1 и Π_2 обслуживаются с помощью адаптивного нециклического алгоритма [5]. Процесс обслуживания каждого потока состоит из 2 этапов с разными интенсивностями. При этом параметры $\mu_{j,1}^{-1}$ и $\mu_{j,2}^{-1}$ задают длительности обслуживания одной заявки на первом и втором этапе соответственно. Также предполагается, что одновременно возможно обслуживание нескольких требований одного потока. Величина $0 < \theta_j \leq 1$ обозначает часть обслуживания, которую необходимо пройти требованию, чтобы можно было начать обслуживать следующую заявку.

В рассматриваемой системе обслуживающее устройство имеет восемь состояний $\{\Gamma^{(1)}, \dots, \Gamma^{(8)}\} = \Gamma$. На следующем рисунке приведен граф смены состояний обслуживающего устройства.



Параметрами системы являются постоянные T_1, T_2, \dots, T_6 , которые определяют длительности состояний $\Gamma^{(1)}, \Gamma^{(2)}, \dots, \Gamma^{(6)}$. Пребывание в состояниях $\Gamma^{(2)}$ и $\Gamma^{(5)}$ может быть продлено на T_2 и T_5 , если очередь по обслуживаемому потоку превышает критическую длину K_1 и K_2 соответственно или на предыдущем продлении пришли новые требования по обслуживаемому потоку. Максимальное количество продлений состояний $\Gamma^{(2)}$ и $\Gamma^{(5)}$ задается константами n_1 и n_2 . В состояниях $\Gamma^{(7)}$ и $\Gamma^{(8)}$ длительности пребывания случайны, зависят от очередности

приход требований и принимают значения на интервалах $(0, T_1]$ и $(0, T_4]$ соответственно. Требования входных потоков поступают в неограниченные накопители O_1 и O_2 . Система рассматривается в дискретные моменты τ_i — моменты смены состояний обслуживающего устройства. Состояние обслуживающего устройства и длины очередей по потокам Π_1 и Π_2 в момент τ_i задаются величинами Γ_i и $(\kappa_{1,i}, \kappa_{2,i}) = \kappa_i$. Число заявок потоков Π_1 и Π_2 , поступивших в систему за промежуток $[\tau_i, \tau_{i+1})$ обозначим через $(\eta_{1,i}, \eta_{2,i}) = \eta_i$. Алгоритм управления учитывает очередьность прихода заявок, поэтому обозначим через η'_i вектор, принимающий значение $y_0 = (0, 0)$, если на i -ом такте $[\tau_i, \tau_{i+1})$ в систему не поступило ни одной заявки, и принимающий значения $y_1 = (1, 0)$ и $y_2 = (0, 1)$, если на i -ом такте первой пришла заявка (или заявки) потоков Π_1 и Π_2 соответственно. Максимальное число заявок потоков Π_1 и Π_2 , которые могут быть обслужены на промежутке $[\tau_i, \tau_{i+1})$, обозначим через $(\xi_{1,i}, \xi_{2,i}) = \xi_i$.

Адаптивный алгоритм смены состояний обслуживающего устройства из множества Γ задается с помощью рекуррентного соотношения:

$$\Gamma_{i+1} = \begin{cases} \Gamma^{(3j-2)}, & \left\{ \left[\Gamma_i = \Gamma^{(3s)} \right] \cap \left[(\kappa_{j,i} > 0) \cup (\kappa_{s,i} \geq K_s) \cup (\eta'_i = y_j) \right] \right\} \cup \\ & \cup \left\{ \left[\Gamma_i = \Gamma^{(3j)} \right] \cap \left[\kappa_{s,i} = 0 \right] \cap \left[\kappa_{j,i} \leq K_j \right] \cap \left[\eta'_i = y_j \right] \right\}, \\ \Gamma^{(3j-1)}, & \left\{ \Gamma_i = \Gamma^{(3j-2)} \right\} \cup \left\{ \left[\Gamma_i = \Gamma^{(6+j)} \right] \cap \left[\eta'_i = y_j \right] \right\}, \\ \Gamma^{(3j)}, & \left\{ \Gamma_i = \Gamma^{(3j-1)} \right\} \cup \left\{ \left[\Gamma_i = \Gamma^{(6+j)} \right] \cap \left[\eta'_i \neq y_j \right] \right\}, \\ \Gamma^{(6+j)}, & \left[\Gamma_i = \Gamma^{(3s)} \right] \cap \left[\kappa_{j,i} = 0 \right] \cap \left[\kappa_{s,i} < K_s \right] \cap \left[\eta'_i = y_0 \right]; \end{cases}$$

при $j, s = 1, 2, j \neq s$.

Динамика длины очереди задается следующими рекуррентными соотношениями

$$\kappa_{j,i+1} = \begin{cases} \max\{0; \kappa_{j,i} + \eta_{j,i} - \xi_{j,i}\}, & \text{если } \Gamma_i \in \Gamma \setminus \{\Gamma^{(3)}, \Gamma^{(6)}\}; \\ \eta_{j,i} + \max\{0; \kappa_{j,i} - \xi_{j,i}\}, & \text{если } \Gamma_i \in \{\Gamma^{(3)}, \Gamma^{(6)}\}. \end{cases}$$

Векторная последовательность $\{(\Gamma_i, \kappa_i); i = 0, 1, \dots\}$ является марковской [5]. Марковская цепь является разложимой. Множество состояний цепи разбивается на класс несущественных состояний и неразложимый апериодическим класс существенных состояний.

Для последовательности $\{(\Gamma_i, \kappa_i); i = 0, 1, \dots\}$ доказаны теоремы существования предельного распределения.

Обозначим через константы $M_1 = (1 + 2\alpha_1 + \alpha_1\beta_1(2/(1 - \gamma_1) + 1/(1 - \gamma_1)^2))p_1$ и $M_2 = (1 + 2\alpha_2 + \alpha_2\beta_2(2/(1 - \gamma_2) + 1/(1 - \gamma_2)^2))p_2$ математические ожидания числа требований для вызывающих моментов потоков Π_1 и Π_2 .

Теорема 1. Для существования предельного распределения марковской последовательности $\{(\Gamma_i, \kappa_i); i \geq 0\}$ необходимо выполнение неравенства

$$\frac{\theta_1 \lambda_1 M_1}{\mu_{1,2}} + \frac{\theta_2 \lambda_2 M_2}{\mu_{2,2}} < 1.$$

Теорема 2. Для существования предельного распределения векторной последовательности $\{(\Gamma_i, \kappa_i); i \geq 0\}$ достаточно выполнения неравенств

$$\lambda_j M_j T - L_j < 0, \quad j = 1, 2,$$

где $T = T_1 + T_3 + T_4 + T_6 + n_1 T_2 + n_2 T_5$ — максимальная длительность основного цикла и L_j — максимальное число обслуженных требований за основной цикл.

3. Численное исследование системы

Аналитически не удается найти такие важные с практической точки зрения характеристики, как среднее время ожидания обслуживания произвольного требования и средние длины очередей по потокам.

Для численного исследования данных характеристик реализована имитационная модель системы адаптивного управления конфликтными потоками неоднородных требований в виде программы, написанной на языке C++. Каждая реализация функционирования системы определяется следующими входными данными:

- 1) параметры $\alpha_j, \beta_j, \gamma_j, \lambda_j$ входных потоков;
- 2) параметры $T_1, T_2, \dots, T_6, \mu_{j,1}, \mu_{j,2}, \theta_j, K_j, n_j$ системы;
- 3) начальные значения $\Gamma^{(r)}, x_1, x_2$ случайных элементов $\Gamma_0, \kappa_{1,0}, \kappa_{2,0}$.

Имитационное моделирование позволяет построить модель более приближенную к реальной системе. При этом способе исследования можно наблюдать за процессами обслуживания в каждый момент времени. Обозначим через $\gamma_j^0(l)$ время ожидания обслуживания заявки с номером $l = 1, 2, \dots$ потока Π_j при нулевых начальных очередях. Величина

$$\bar{\gamma}_j^0(n) = \frac{1}{n} \sum_{l=1}^n \gamma_j^0(l)$$

определяет выборочное среднее время ожидания обслуживания в системе первых n требований потока Π_j . Оценку $\bar{\gamma}^*$ среднего времени ожидания обслуживания произвольного требования будем вычислять по формуле среднего взвешенного

$$\bar{\gamma}^* = \frac{\lambda_1 M_1 \bar{\gamma}_1^0 + \lambda_2 M_2 \bar{\gamma}_2^0}{\lambda_1 M_1 + \lambda_2 M_2}.$$

Выборочные средние времена ожидания обслуживания будем вычислять для системы, в которой достигнут стационарный режим. Момент окончания переходного процесса определяется с помощью алгоритма, описанного в [6].

Приведем пример поиска квазиоптимальных значений параметров при следующем начальном наборе входных параметров:

- 1) параметры входных потоков $\alpha_1 = 0.842, \beta_1 = 0.655, \gamma_1 = 0.448, \lambda_1 = 0.03, \alpha_2 = 0.842, \beta_2 = 0.655, \gamma_2 = 0.448, \lambda_2 = 0.05$;
- 2) параметры системы обслуживания $T_1 = 4, T_2 = 5, T_3 = 4, T_4 = 4, T_5 = 5, T_6 = 4, \mu_{1,1} = 0.25, \mu_{2,1} = 0.25, \mu_{1,2} = 0.5, \mu_{2,2} = 0.5, \theta_1 = 0.5, \theta_2 = 0.5, K_1 = 10, K_2 = 10, n_1 = 10, n_2 = 10$;
- 3) начальное состояние системы $\Gamma^{(r)} = \Gamma^{(1)}, x_1 = 0, x_2 = 0$.

Параметры входного потока соответствуют разбиению транспортного потока Бартлетта, приведенному в [7]. При этом, для входных потоков принято одинаковые распределения размеров групп требований, но разные интенсивности λ_1 и λ_2 потоков групп (потоков вызывающих моментов). В силу физических ограничений для оптимизации доступны только следующие параметры $T_2, T_5, n_1, n_2, K_1, K_2$. Оптимизация производится поэтапно по парам параметров (T_2, T_5) , (n_1, n_2) и (K_1, K_2) путем сокращенного перебора. В качестве критерия эффективности функционирования системы будем использовать оценку $\bar{\gamma}$ среднего времени ожидания обслуживания произвольного требования.

Используя алгоритм сокращенного перебора, были получены квазиоптимальные параметры $T_2^* = 5, T_5^* = 6, n_1^* = 3, n_2^* = 9, K_1^* = 1, K_2^* = 11$, которым соответствуют оценки $\bar{\gamma}_1^* = 8.0009, \bar{\gamma}_2^* = 7.5249, \bar{\gamma}^* = 7.7034$ средних времен ожидания обслуживания.

В табл. 1 приведены шаги алгоритма оптимизации.

T_2	T_5	n_1	n_2	K_1	K_2	$\bar{\gamma}_1^*$	$\bar{\gamma}_2^*$	$\bar{\gamma}^*$
5	5	10	10	10	10	18.998	27.859	25.536
5	5	10	10	1	9	7.8603	7.6275	7.7148
5	5	3	9	1	9	7.8763	7.6168	7.7141
5	6	3	9	1	9	8.0119	7.5446	7.7198
5	6	3	9	1	11	8.0009	7.5249	7.7034
5	6	9	6	1	11	7.9783	7.5437	7.7067

Таблица 1. Поиск квазиоптимальных параметров системы управления

Из таблицы следует, что наибольшее влияние на уменьшение оценки среднего времени ожидания обслуживания повлияло уменьшение критической длины очереди для менее интенсивного потока Π_1 на первом шаге алгоритма.

4. Заключение

Для описанной в работе системы приведены теоремы о существовании предельного распределения. Для изучения среднего времени ожидания обслуживания было использовано имитационное моделирование. Данный метод позволил получить оценки для характеристик функционирования системы, которые не удается вычислить аналитически.

Приведен алгоритм поиска квазиоптимальных параметров для системы, в которой параметры входных потоков получены при разбиении на группы реального транспортного потока.

ЛИТЕРАТУРА

1. Федоткин М. А., Федоткин А. М., Кудрявцев Е. В. Динамические модели неоднородного потока транспорта на магистралях // Автоматика и телемеханика. 2020, № 8, с. 149–164.
2. Кокс Д., Льюис П. Статистический анализ последовательностей событий. М.: Мир, 1969, 312 с.
3. Федоткин М. А., Кудрявцев Е. В. Компьютерная обработка статистических данных потока катастроф на угольных шахтах // Материалы Межрегиональной научно-практической конференции «Статистика в современном обществе: её роль и значение в вопросах государственного управления и общественного развития». Н.Новгород, 2015, с. 451–457.
4. Федоткин М. А., Кудрявцев Е. В. Изучение характеристик транспортного потока большой плотности. Деп. в ВИНТИ 14.01.2014, №14, с. 33.
5. Федоткин М. А., Кудрявцев Е. В. Анализ дискретной модели системы адаптивного управления конфликтными неоднородными потоками // Вестник московского университета. Серия 15: Вычислительная математика и кибернетика. 2019, № 1. с. 19–26.
6. Федоткин М. А., Кудрявцев Е. В. Исследование переходного процесса адаптивного управления потоками неоднородных требований путем имитационного моделирования // Сборник материалов XVIII Международной конференции «Информационные технологии и математическое моделирование (ИТММ-2019)». Томск: НТЛ. 2019, Часть 2, с. 207–212.
7. Федоткин М. А., Кудрявцев Е. В. Оценка параметров вероятностной модели интенсивного транспортного потока // В сборнике: «Распределенные компьютерные и коммуникационные сети: управление, вычисление, связь (DCCN-2013)». 2013, с. 365–372.

UDC: 519.21

On asymptotic analysis of quasi-regenerative processes

G.A. Zverkina¹

¹V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 65
Profsoyuznaya street, Moscow 117997, Russia

Abstract

In queuing theory and related fields, many methods of asymptotic analysis of the behaviour of stochastic processes are based on the analysis of regenerative processes. However, the behaviour of some complex technical systems cannot be described by regenerative processes.

For example, the behaviour of one part of some system affects the behaviour of another part of this system, and vice versa. In this case, the total process of behaviour of both subsystems is not regenerative.

However, in many technical systems, these non-regenerative processes are close “in some sense” to regenerative processes. Hence, its asymptotic behaviour can be studied.

In this paper, we give a definition of a process close to regenerative and propose a method for studying its asymptotic behaviour.

Keywords: regenerative process, Markov process, embedded renewal process, quasi-regenerative process, quasi-renewal process, convergence rate, total variation metric

1. Introduction

As is well known, many processes in queuing theory and in related areas are described by regenerative processes. However, in some situations the processes are non-regenerative, but in some sense they are close to regenerative processes.

First of all, we will recall the definition of regenerative process (see, e.g., [1]), then we will give examples of non-regenerative processes already studied, and then move on to the definition of quasi-regenerative processes and some of their properties.

Definition 1. A random process is called regenerative if there exists an increasing sequence $\{t_i\}_{i=0,1,2,\dots}$, such, that the random elements $\Theta_i \stackrel{\text{def}}{=} \{X_t, t \in [t_{i-1}; t_i]\}$ are i.i.d. $\forall i = 1, 2, \dots$

Times t_i are named *regeneration times*.

Denote $\tau_i \stackrel{\text{def}}{=} t_{i+1} - t_i$, and let \mathcal{P}_t be a distribution of regenerative process at the time t . \triangleright

For regenerative process, if $\mathbb{E} \tau_i < \infty$, then the process is ergodic, i.e. it has the limit invariant distribution \mathcal{P} ; $\mathcal{P}_t \Rightarrow \mathcal{P}$. In the case when $\mathbb{E} (\tau_i)^k < \infty$, $k > 1$, the convergence rate of $\mathcal{P}_t \Rightarrow \mathcal{P}$ can be estimated using Lorden's inequality – see [8].

The times $t_0, t_1, t_2, \dots, t_n, \dots$ form an embedded renewal process for the regenerative process; here it is useful to recall the definition of the renewal process (see, e.g., [5])

Definition 2. The renewal process N_t is a counting process:

$$N_t \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} \mathbf{1} \left\{ \sum_{s=1}^i \xi_s \leq t \right\},$$

where $\{\xi_1, \xi_2, \dots\}$ – i.i.d. positive random variables with d.f. $F(s)$. The value of N_t changes at the times $t_k = S_k \stackrel{\text{def}}{=} \sum_{j=1}^s \xi_j$. These moments t_k are called renewal moments.

For this renewal process the random variables $B_t \stackrel{\text{def}}{=} t - S_{N_t}$ and $W_t \stackrel{\text{def}}{=} S_{N_t+1} - t$ are backward renewal time and forward renewal time accordingly. \triangleright

Usually, queuing systems and reliability systems are described by regenerative processes. However, in some cases in applied problems, the stochastic process (or a component of the stochastic process) can be non-regenerative.

Example 1. Consider the system which consists of the main element and redundant element with reliability characteristics that depend on each other and on the time.

So, the intensities of failure and repair of both elements are the functions of their states, or the *full system state*. This *full state of the system* is described by pairs: $X_t \stackrel{\text{def}}{=} ((i_t, x_t), (j_t, y_t))$, where $i_t = 0$ or $i_t = 1$ if the first (main) element is working or not at the time t correspondingly. And the value x_t is equal to the time elapsed from the last change in the state of i_t of the first (main) element to the time t . The pair (j_t, y_t) describes the state of the reserve element at the time t in the same way.

Suppose that the intensities of failure and repair are the function $\lambda_k(X_t)$ and $\mu_k(X_t)$ respectively (k is the number of element).

The process X_t is non-regenerative, its ergodicity is proved in [6, 7]. Moreover, for the case of bounded and separated from zero intensities $\lambda_k(X_t)$ and $\mu_k(X_t)$, the convergence rate of the distribution of X_t is estimated in [7]. \triangleright

In [3, 4, 9] the notion of quasi-renewal (or generalized renewal) process has been defined.

Definition 3. Let

- 1) $\xi_j = \min\{\zeta_j; \eta_j\}$, where $\{\zeta_j\}$ are i.i.d. random variables defined by (generalized) intensities $\varphi_i(s) \equiv \varphi(s)$, and $\zeta_i \perp\!\!\!\perp \eta_j$ for all i, j ; η_j are defined by (generalized) intensities $\mu_j(s)$;
- 2) There exists some (generalized) measurable function $Q(s)$ such that for all $s \geq 0$, $\varphi(s) + \mu_j(s) = \lambda_i(s) \leq Q(s)$;
- 3) $\int_0^\infty \varphi(s) ds = \infty$, and $\int_0^\infty \left(x^{k-1} \exp \left(- \int_0^x \varphi(s) ds \right) \right) dx < \infty$ for some $k \geq 2$;
- 4) $Q(s)$ is bounded in some neighbourhood of zero;
- 5) $\varphi(s) > 0$ a.s. for $s > T \geq 0$.

If conditions 1–4 are satisfied, then the counting process

$$N_t \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} \mathbf{1} \left\{ \sum_{k=1}^i \xi_k \leq t \right\} \quad (1)$$

is named *quasi-renewal process (or generalized renewal process)*.

For this quasi-renewal process the random variables $B_t \stackrel{\text{def}}{=} t - S_{N_t}$ and $W_t \stackrel{\text{def}}{=} S_{N_t+1} - t$ are backward renewal time and forward renewal time accordingly. \triangleright

Remark 1. In Definition 3, the assumption 1 holds: the random variables ξ_i and ξ_j are dependent, and this dependency is “weak” dependency in some sens.

Also the intensity of renewal time of considered quasi-renewal process is a sum of two independent processes: one of them is a classic renewal process (with renewal times ζ_i) and quasi-renewal process with (dependent) renewal times η_i .

Such a process describes some process that has some minimal intensity and additional increase in intensity due to the influence of some external factors. \triangleright

Remark 2. In Definition 3, the assumptions 3 and 4 hold: $\mathbf{E} \xi_i > 0$, $\text{Var} \xi_i^2 > 0$.

\triangleright

Remark 3. In Definition 3, the assumptions 1 and 2 hold:

$$F_i(t) = \mathbf{P}\{\xi_i \leq t\} = 1 - \exp \left(\int_0^t -\lambda_i(s) ds \right) \geq 1 - \exp \left(\int_0^t -Q(s) ds \right)$$

$$\Rightarrow \exists \mathbf{E} \xi_i^2 < \infty.$$

\triangleright

Remark 4. In Definition 3, the assumption 5 reports that the renewal process under study is a delay renewal process, and a delay time does not exceed T . \triangleright

Example 2. Consider the process $X_t \stackrel{\text{def}}{=} (B_t^{(1)}, B_t^{(2)}, B_t^{(3)}, \dots, X_t^{(m)})$, where the processes $B_t^{(i)}$, $i = 1, 2, \dots, m$ are the backward renewal times of *dependent* quasi-renewal processes $N_t^{(i)}$, $i = 1, 2, \dots, m$.

At the time t , the state of the process $B_t^{(i)}$ is $B_t^{(i)} = t - \sum_{j=1}^{N_t^{(i)}} \xi_i^{(j)}$ – elapsed renewal time or backward renewal time of process $N_t^{(i)}$.

The intensities of the processes $N_t^{(i)}$, $i = 1, 2, \dots, m$ are $\varphi(B_t^{(i)}) + \mu_i(X_t)$.

This is a Generalized Markov Modulated Poisson Process (GMMPP), invented and studied in [9]. In [9], the ergodicity of this non-regenerative process is proved, and the method to obtain the bounds for its convergence rate is proposed. \triangleright

In both these examples, the proof of ergodicity is based on the coupling method and generalized Lorden's inequality ([3, 4]).

2. Quasi-regenerative processes

The examples considered above lead to the concept of a quasi-regenerative process.

Definition 4. Markov random process X_t is called quasi-regenerative if there exists an other *regenerative* Markov random process \tilde{X}_t such that at all times $t \geq 0$, the distributions of X_t and \tilde{X}_t are equal: $\tilde{X}_t \stackrel{D}{=} X_t$ for all $t \geq 0$. \triangleright

Remark 5. All regenerative process can be named quasi-regenerative process.

Quasi-regenerative process which is non-regenerative, can not only one embedded renewal process; as a general rule, a quasi-regenerative process has two or more embedded dependent quasi-renewal processes. \triangleright

In many situations, a non-regenerative processes in queueing theory and related fields are composed of two or more dependent quasi-renewal processes (or alternating quasi-renewal processes).

The Definition 3 of quasi-renewal processes implies the existence of a finite upper bound for the mathematical expectation of the length of the quasi-renewal periods.

Lemma 1. The quasi-regenerative process is ergodic.

The proof of Lemma 1 is based on the coupling method. More precisely, for quasi-regenerative process X_t the modification of coupling method “*successful coupling*” used for construction of the regenerative “copy” $\tilde{X}_t \stackrel{D}{=} X_t$ (see, e.g., [2, 7, 9]). I.e. Thus, the regenerative period of the regenerative “copy” $\tilde{X}_t \stackrel{D}{=} X_t$ is a geometric sum of the quasi-renewal periods of the components of the process X_t .

This consideration implies the technology for obtaining upper bounds for convergence rate of quasi-regenerative process distribution to the stationary one.

Theorem 1. If the k -th moments of all embedded quasi-renewal processes of quasi-regenerative processes X_t are less than some number $K < \infty$, then there exists

calculated constant C such that the distance between the distribution of X_t and its stationary distribution is less than $\frac{C}{t^{k-1}}$ for all $t > 0$. \triangleright

Theorem 2. If the α -th exponential moments of all embedded quasi-renewal processes of quasi-regenerative processes X_t are less than some number $K < \infty$, then there exists calculated constant C , and calculated $\beta \in (0, \alpha)$, such that for all $\gamma \in (0, \beta)$ the distance between the distribution of X_t and its stationary distribution is less than $C \exp(-\gamma t)$ for all $t > 0$. \triangleright

The proof of both these Theorems is similar to the one in [9]. The upper bounds for convergence rate in any situations can be improved by the use of the particular qualities of the distributions of the components of quasi-regenerative processes.

In this short paper, a full presentation of the proofs is impossible. The full proofs will be given in an extended publication.

3. Conclusion

Quasi-regenerative process describes the behaviour of a complex technical system which consists of many dependent subsystems functioning simultaneously.

The study and use of quasi-regenerative processes will give a significant contribution to the development of technology for obtaining convergence rate estimates for various stochastic models.

Acknowledgments

The work is supported by RFBR, project No 20-01-00575 A. The author is grateful to anonymous reviewers for their valuable comments.

REFERENCES

1. Afanasyeva, L. G., Tkachenko A. V., Multichannel queueing systems with regenerative input flow // Theory Probab. Appl., 58:2 (2014), 174–192
2. Griffeath, D. A maximal coupling for Markov chains // Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete – 1975 – Volume 31 – Issue 2, P. 95–106.
3. Kalimulina E. Yu., Zverkina G. A. On some generalization of Lorden's inequality for renewal processes // arXiv:1910.03381v2, P.1–8.
4. Kalimulina E. Yu., Zverkina G. A. On generalized intensity function and its application to the backward renewal time estimation for renewal processes //

- Proceedings of the 5th International Conference on Stochastic Methods (ICSM-5, 2020). M.: RUDN, 2020. P. 306–310.
5. Smith, W. L., Renewal theory and its ramifications // J. Roy. Statist. Soc. Ser. B, 20:2 (1958), 243-302
 6. Veretennikov, A. On Polynomial Recurrence for Reliability System with a Warm Reserve, Markov Processes and Related Fields. 2019. Vol. 25. P. 745–761. (DOI 10.1007/s11134-019-09626-x)
 7. Zverkina, G. A. System with Warm Standby// Computer Networks 26th International Conference, CN 2019, Kamień Śląski, Poland, June 25–27, 2019, Proceedings. P. 387–399.
 8. Zverkina, G. A. Lorden's inequality and coupling method for backward renewal process / Proceedings of the 20th International Scientific Conference “Distributed Computer and Telecommunication Networks: Control, Computing, Communication” (DCCN-2017, Moscow). M .: Tekhnosfera, 2017.P. 484–491.
 9. Zverkina, G. A. Ergodicity and Polynomial Convergence Rate of Generalized Markov Modulated Poisson Processes // Proceedings of the 23rd International Conference on Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN-2020, Moscow). Cham: Springer, 2021. Vol.1337. P. 367–381.

УДК: 519.248

О предельном распределении максимума стационарного времени ожидания в GI/G/1 с Экспоненциальным-Парето обслуживанием

И.В. Пешкова^{1,2}

¹Петрозаводский государственный университет, пр. Ленина, 33, Петрозаводск, Россия

²Институт прикладных математических исследований Карельского научного центра РАН, ул. Пушкинская, 11, Петрозаводск, Россия

iaminova@petrsu.ru

Аннотация

В данной работе мы изучаем экстремальное поведение стационарного времени ожидания в системах $GI/G/1$, в которых время обслуживания имеет распределение, заданное смесью Экспоненциального распределения и Парето распределения II типа. Распределение такого типа принадлежит к специальному подклассу субэкспоненциальных распределений, что позволяет применить известную аппроксимацию хвоста функции распределения времени ожидания через равновесное распределение времени обслуживания. Найдены нормализующие последовательности, при которых предельное распределение максимума стационарных времен ожидания в $GI/G/1$ имеет распределение Фреше.

Ключевые слова: системы обслуживания, экстремальные значения, Экспоненциальное -Парето распределение

1. Введение

Смеси распределений широко используются для представления неоднородных данных. Конечная смесь экспоненциального и Парето распределений представляет особый интерес [4] для моделирования времени обработки данных (связанным со временем обслуживания в моделях массового обслуживания), поскольку аппроксимирует смесь распределений с так называемыми *легкими и тяжелыми хвостами*. Например, телекоммуникационные устройства могут создавать огромные объемы трафика данных, например мультимедийного контента, и время обработки может сопровождаться "всплесками" задержек. Напротив, показания сенсоров (значения температуры, давления или освещенности, передаваемые датчиками) могут создавать «малые и скачкообразные» объемы

данных, не требующие длительной обработки. В таких случаях модель конечной Экспоненциальной-Парето смеси может хорошо работать.

Мы рассматриваем систему обслуживания $GI/G/1$ с Экспоненциальным-Парето распределением времени обслуживания. Используя известный критерий [2] в статье показано, что это распределение принадлежит к специальному подклассу \mathcal{S}^* субэкспоненциальных распределений. В этом случае асимптотика хвоста функции распределения стационарного времени ожидания определяется равновесным распределением времени обслуживания [1]. Эта асимптотика применяется для получения предельного распределения максимума времени ожидания на основе теории экстремальных значений. Для реализации этой идеи были предложены выражения для нормализующих констант, что в итоге позволило получить, что предельное распределение стационарного времени ожидания имеет распределение Фреше.

2. Асимптотика распределения стационарного времени ожидания

Рассмотрим систему массового обслуживания $GI/G/1$ с независимыми и одинаково распределенными (н.о.р.) интервалами между приходами заявок и н.о.р. временами обслуживания. Дисциплина обслуживания – первым пришел, первым обслужился. Пусть B – функция распределения (ф.р.) времени обслуживания S с конечным математическим ожиданием $ES = 1/\mu$. Известно, что если B принадлежит к классу так называемых *субэкспоненциальных распределений*, то стационарное распределение времени ожидания W имеет следующую асимптотику [1]:

$$\mathbf{P}(W > x) \sim \frac{\rho}{1 - \rho} \mathbf{P}(S_e > x), \quad x \rightarrow \infty, \quad (1)$$

где равновесное время обслуживания S_e имеет равновесную плотность $\bar{B}(x)/ES$ и также принадлежит к субэкспоненциальным распределениям (отношение $a \sim b$ в (1) означает асимптотическую эквивалентность, то есть $a/b \rightarrow 1$), ρ – коэффициент загрузки системы.

Для проверки того, что ф.р. B и соответствующее равновесное распределение B_e принадлежат к классу субэкспоненциальных распределений, достаточно проверить, что B принадлежит к специальному подклассу \mathcal{S}^* субэкспоненциальных распределений.

Обозначим функцию интенсивности отказов

$$r_B(x) = \frac{f_B(x)}{\bar{B}(x)},$$

где f_B – плотность распределения, соответствующая ф.р. B . Проверим следующий критерий принадлежности распределения к \mathcal{S}^* .

Если

$$\lim_{x \rightarrow \infty} r_B(x) = 0 \quad \text{и} \quad \lim_{x \rightarrow \infty} xr_B(x) < \infty, \quad (2)$$

то $B \in \mathcal{S}^*$ [2].

Пусть случайная величина (с.в.) S имеет *Экспоненциальное-Парето распределение* [4] с хвостом функции распределения вида:

$$\bar{B}(x) = pe^{-\lambda x} + (1-p) \left(\frac{x_0}{x_0 + x} \right)^\alpha, \quad \lambda > 0, \quad \alpha > 0, \quad x_0 > 0, \quad x \geq 0, \quad (3)$$

где $0 < p < 1$ параметр пропорции смеси. Уравнение (3) показывает, что с.в. S совпадает с экспоненциальным распределением с вероятностью p и с распределением Парето с вероятностью $1 - p$.

Выражения для среднего времени обслуживания и равновесного распределения B_e времени обслуживания имеют вид:

$$1/\mu = \frac{p}{\lambda} + \frac{(1-p)x_0}{\alpha - 1}, \quad (4)$$

$$B_e(x) = \mu \int_0^x \bar{B}(t) dt = 1 - \mu \left(\frac{pe^{-\lambda x}}{\lambda} + \frac{(1-p)x_0^\alpha}{(\alpha - 1)(x_0 + x)^{\alpha-1}} \right). \quad (5)$$

Заметим, что оба выражения (4) и (5) существуют при параметре $\alpha > 1$.

Функция интенсивности отказов Экспоненциального-Парето распределения вычисляется по формуле:

$$r_B(x) = \frac{p \lambda a(x) + (1-p) \alpha / (x_0 + x)}{p a(x) + (1-p)}, \quad (6)$$

где

$$a(x) = e^{-\lambda x} \left(1 + \frac{x}{x_0} \right)^\alpha.$$

Легко проверить, что условия (2) выполнены, поскольку

$$r_B(x) \rightarrow 0 \quad \text{при} \quad x \rightarrow \infty \quad \text{и} \quad xr_B(x) \rightarrow \alpha, \quad \text{при} \quad x \rightarrow \infty$$

и, следовательно, Экспоненциальное-Парето распределение принадлежит к подклассу \mathcal{S}^* , что позволяет применить асимптотику (1) для стационарного времени ожидания в рассматриваемой системе.

Обозначим через λ_τ интенсивность входного потока. Вычислим интенсивность загрузки системы

$$\rho = \lambda_\tau \left[\frac{p}{\lambda} + \frac{(1-p)x_0}{\alpha - 1} \right], \quad (7)$$

и подставим (7) и (5) в (1). Получим следующее асимптотическое выражение для хвоста функции распределения стационарного времени ожидания:

$$\mathbf{P}(W > x) \sim \frac{\mu\lambda_\tau}{\mu - \lambda_\tau} \left(\frac{pe^{-\lambda x}}{\lambda} + \frac{(1-p)x_0^\alpha}{(\alpha-1)(x_0+x)^{\alpha-1}} \right) \text{ при } \rightarrow \infty. \quad (8)$$

3. Экстремальное поведение стационарных времен ожидания

Пусть $\{X_n, n \geq 1\}$ — семейство н.о.р. с.в. с ф.р. F . Обозначим $M_n = \max(X_1, \dots, X_n)$ — максимальное значение в последовательности. Очевидно, что имеет место следующее равенство:

$$\mathbf{P}(M_n \leq x) = F^n(x).$$

Известно [3], что если для некоторых последовательностей констант $b_n, a_n > 0, n \geq 1$ нормированный максимум $(M_n - b_n)/a_n$ имеет невырожденную предельную функцию распределения $G(x)$,

$$\mathbf{P}((M_n - b_n)/a_n \leq x) \rightarrow G(x), \quad n \rightarrow \infty, \quad (9)$$

то $G(x)$ принадлежит к одному из трех типов распределений *экстремальных значений*: Гумбеля, Фреше или Вейбулла.

Предположим, что существует последовательность вещественных констант $\{u_n, n \geq 1\}$ такая, что для некоторого $0 \leq \tau \leq \infty$

$$n\bar{F}(u_n) \rightarrow \tau \text{ при } n \rightarrow \infty. \quad (10)$$

Тогда из [3] следует, что

$$\mathbf{P}(M_n \leq u_n) \rightarrow e^{-\tau} \text{ при } n \rightarrow \infty. \quad (11)$$

Если выполнено условие (9), то сходимость (11) сохраняется для любой линейной нормализующей последовательности $u_n(x) = a_n x + b_n, n \geq 1$ и выражение (11) принимает форму

$$\mathbf{P}(M_n \leq u_n(x)) \rightarrow \tau(x),$$

где конкретный вид функции $\tau(x)$ зависит от вида распределения G .

Обозначим $W_n^* = \max(W_1, \dots, W_n)$ — максимальное стационарное время ожидания в последовательности из n с.в., $n \geq 1$.

Лемма 1. *Если время обслуживания в стационарной ($\rho < 1$) системе обслуживания GI/G/1 имеет Экспоненциально-Парето-распределение (3) с параметром $\alpha > 1$, то предельное распределение W_n^* является распределением Фреше, а именно*

$$\mathbf{P}(W_n^* \leq u_n(x)) \rightarrow e^{-\frac{(1-p)\mu\lambda_\tau}{(\alpha-1)(\mu-\lambda_\tau)}x^{1-\alpha}} \quad \text{при } n \rightarrow \infty, \quad (12)$$

где нормализующие последовательности имеют вид

$$u_n(x) = a_n x + b_n = x_0^{\alpha/(\alpha-1)} x n^{1/(\alpha-1)} - x_0. \quad (13)$$

4. Заключение

В статье получено предельное распределение максимума стационарного времени ожидания для системы $GI/G/1$ с Экспоненциальным-Парето распределением времени обслуживания. Этот результат может быть использован для исследования экстремальных значений стационарного времени ожидания в системах с неоднородным потоком обслуживания, аппроксимируемым смесью распределений с "легкими" и тяжелыми" хвостами.

Литература

1. Asmussen S., Kluppelberg C., Sigman K. Sampling at subexponential times with queueing applications // Report TUM M9804. 1998.
2. Goldie C. M., Kluppelberg C. A // Adler, Robert J. and Feldman, Raisa E. and Taqqu, Murad S. (eds) A practical Guide to Heavy Tails: Statistical Techniques for Applications. Birkhauser Boston Inc., 1998.
3. Leadbetter M.R., Lindgren G., Rootzin H. Extremes and Related Properties of Random Sequences and Processes. Springer, New York, 1983. <https://doi.org/10.1007/978-1-4612-5449-2>
4. Peshkova, I. Morozov E. and Maltseva M. On comparison of multiserver systems with Exponential-Pareto mixture distribution // Communications in Computer and Information Science, Springer, 2020. V. 1231, P 141–152. https://doi.org/10.1007/978-3-030-50719-0_11
5. Peshkova I., Morozov E. and Maltseva M. On regenerative estimation of extremal index in queueing systems // Vishnevskiy V.M., Samouylov K.E., Kozyrev D.V. (eds) Distributed Computer and Communication Networks: Control, Computation, Communications. DCCN 2021. Lecture Notes in Computer Science, Springer, Cham, 2021. V. 13144, P 251-264. https://doi.org/10.1007/978-3-030-92507-9_21

UDC: 519.876.5

Comparison of approaches to component reliability allocation for distributed control systems

Aleksandr Moshnikov¹

¹ITMO University, 49 Kronverksky Pr., St. Petersburg, Russian Federation

moshnikov.alex@gmail.com

Abstract

Various approaches to reliability allocation of elements are considered. A comparison of various different methods for solving the CAP problem for a system with a network structure is given. The report describes a mathematical model of the reliability of a distributed control system, statistical modeling is used to determine the parameters of reliability and importance metric of elements.

Keywords: Monte Carlo method, Reliability estimation, Control systems design, Reliability allocation

1. Introduction

As production develops, equipment management systems become more complex. New hardware and software are emerging, and the architecture of systems is becoming more and more complex. Control systems of modern technological equipment are built on the principle of vertically integrated multi-level networks. When designing such networks, the issue of reliability and safety is acute [1].

Designing highly reliable systems requires a systematic approach and working with requirements at the stage of shaping the appearance of the product being developed. To ensure the best design solution, the problem of optimizing design reliability is solved. At the moment, several such typical tasks have been formed, namely[2]:

- redundancy allocation problem (RAP), aims to achieve the maximum system reliability by allocating the redundant components suitably with the cost constraints;
- reliability-RAP (RRAP) is a type of system reliability optimization for obtaining the optimal solution with the highest system reliability by adjusting the component reliability and redundancy with consideration of the limitations (reliability, cost, weight, and volume);

- component assignment problem (CAP), aims to generate optimal assignment with maximum system reliability by assigning n available components into n positions;
- complex reliability problem (CRP), difficult system ROP for complex systems with large-scale or complicated tasks to maximize system performance with limited resources.

2. Optimization methods for CAP

2.1. Reliability model. It is natural to assume that different elements have different effects on the behavior of the system in terms of reliability. Quantifying the nature of the impact of elements on the behavior of the system is of particular importance when analyzing systems. This makes it possible to identify the weaknesses of the system, choose the optimal redundancy, and rationally influence the reliability of the system as a whole. For this purpose, special characteristics called importance and contribution are proposed.

The probability of failure of element can be determined by $q(t) = e^{-\lambda t}$, where λ is the equipment failure rate. System probability of uptime is $R(\pi) = h(p)$, where $h(p)$ structural reliability function.

The importance of the element e_i in the system is defined as a private derivative of the availability factor (the probability of) the system availability (the probability of) the element for which an analysis of its importance:

$$I_{BIM}(i, p) = \frac{\partial h(p)}{\partial p_i} \quad (1)$$

This characteristic is called Birnbaum importance (BIM). The importance is estimated by the number of times the system availability coefficient increases when the element availability coefficient increases. BIM-importance does not depend on the readiness coefficient of the element p , but depends only on p_j for all $i = j$ in satisfies the inequalities $0 \leq I_{BIM}(i) \leq 1$.

Birnbaum's importance metric was applied to develop a heuristic for CAP, trying to assign more reliable components to positions with larger BIM values.

In addition to the Birnbaum metric, other metrics have become widespread in reliability and safety analyses.

2.2. Methods overview. The CAP is to find the optimal permutation of π at which the reliability of the system is maximal. CAP is an NP-hard problem because of its combinatorial nature. When the optimal arrangement depends only on the order of the reliability components, this is called an invariant optimal arrangement, since it exists independently of the reliability values of the components.

In addition to the Birnbaum metric, other metrics have become widespread in reliability and safety analyses. Kontoleon proposed the heuristic [3], in which the basic logic is the monotone property of the BIM-reliability importance. Maximum systems reliability is expected to be achieved by gradually improving the reliability of the position (i.e., by assigning a component of higher reliability) that has the largest BIM-reliability importance value.

Heuristics were further developed in research [4] and [5].

The table 1 shows an algorithm for solving the CAP by a group of ZK heuristics.

Step	Description
<i>I</i>	Generate an initial arrangement randomly, $\pi = (\pi_1, \dots, \pi_n)$
<i>II</i>	Calculate I_{BIM} for all positions from position 1 to position n
<i>III</i>	for $k = 1$ to $n - 1$, do the loop Find positions m and r such that $(\pi_m) = k$ and $\pi_r = k + 1$ If $(I_{BIM})(m) > (I_{BIM})(r)$ and $R(P, \pi) > R(P, \pi(m, r))$ exchange the assignments of components (π_m) and π_r
<i>IV</i>	If there is no exchange in Step III, output the final assignment; otherwise, go to Step II

Table 1. ZKA algorithm

The table 2 shows an algorithm for solving the CAP by a group of LK heuristics.

Step	Description
<i>I</i>	Assign component 1 to all positions that are set $P_{i1} = P_{11}$ and $\pi_i = 1$ for $i = 1, 2, \dots, n$
<i>II</i>	Let $S = 1, 2, \dots, n$, which is the set of available positions
<i>III</i>	for $k = n$ to 2, do the loop Calculate $I_{BIM}(i)$ for all $i \in S$ Find $m \in S$, which meets that $I_{BIM}(m) = \max_{i \in S} I_{BIM}(i)$ Let $S = S / m$, assign component k to position m
<i>IV</i>	If there are no components in S , output the final assignment otherwise, go to Step II

Table 2. LKA algorithm

In addition to the two basic heuristics, there are 6 modifications of them, using different ways of reassigning and selecting the initial set of component distributions.

3. Distributed control system model

3.1. Reliability model of DCS. Distributed process control system (DCS) is a group of technical and software solutions designed to automate technological processes in industrial enterprises. As a rule, automatic process control systems are understood as a complete solution that ensures the automated execution of the basic operations of the technological process of production. The representation of DCS in the form of a graph is shown in Figure 1.

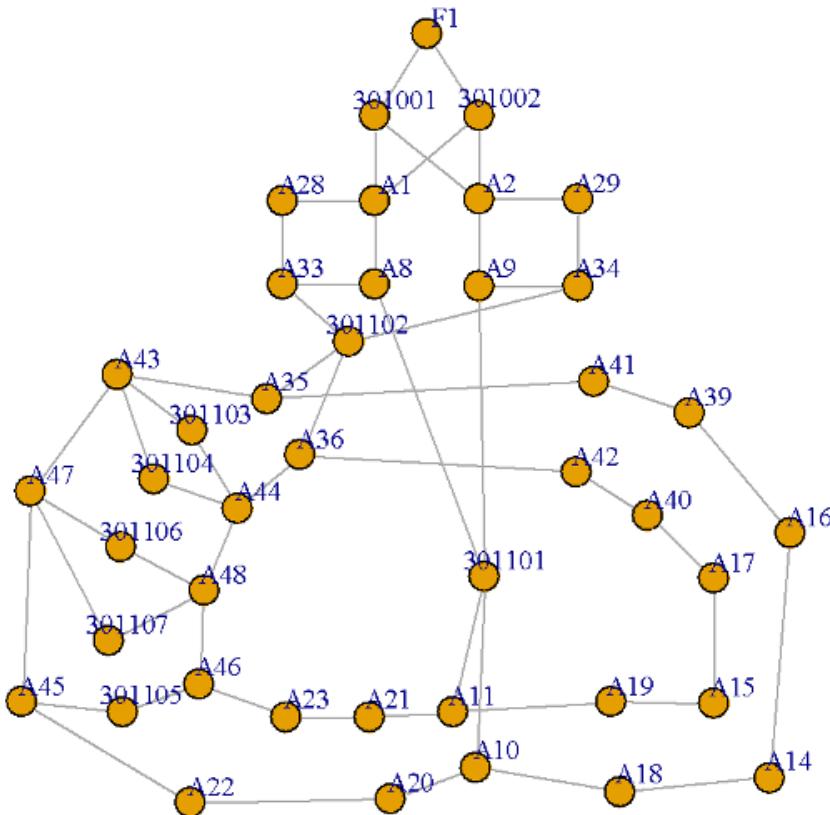


Fig. 1. Distributed control system graph model

3.2. Monte-Carlo method for reliability estimating. The need to obtain reliability parameters of a complex system with a network structure limits the possibility of using classical methods of reliability analysis, like reliability block diagrams or failure trees. In fact, the only way to obtain estimates of uptime

probability is to use statistical modeling (Monte Carlo method) with an acceptable number of iterations, it is possible to obtain accurate results [6] and [7]. .

3.3. Program realisation and experiment. To compare various heuristics, a software module was developed that provides the calculation of system reliability characteristics and BIM metric. The R programming language was used for implementation. R is a programming language for statistical data processing and working with graphics, as well as a free open-source computing environment for the GNU project. The R language contains tools that allow you to create multiple parallel threads of calculations (due to simultaneous loading of several processor cores) and reduce the time spent on modeling several times.

As an example, 41 vertex graphs and data on the reliability of elements that can act as vertices were used to test the efficiency and accuracy of the heuristics of the CAP solution. The probability of vertex uptime was assumed in the range from 0.8 to 0.99.

4. Conclusion

A study of heuristics for solving the CAP problem was conducted. A distributed control system with a network structure was chosen as an example. To obtain estimates of the reliability of the system, the Monte Carlo method was used.

ZKD heuristic performs best for the ones with high reliable or arbitrary components. In addition, the ZKB and ZKD heuristics, which use the B-reliability importance to effectively find the pairs of components for exchange, are superior to the ZKA and ZKC heuristics.

REFERENCES

1. Zuo, Ming Jian and Way Kuo. “Optimal Reliability Modeling: Principles and Applications.” 2002.
2. Si S., Zhao J., Cai Z, Dui H. (2020). Recent advances in system reliability optimization driven by importance measures. *Frontiers of Engineering Management*. 7. 1-24. 10.1007/s42524-020-0112-6.
3. Kontoleon J.M. Optimal link allocation of fixed topology networks. *IEEE Transactions on Reliability* 28, 145–147. 1979.
4. Lin F, Kuo W 2002 Reliability importance and invariant optimal allocation. *Journal of Heuristics* 8: 155-172.
5. Kuo, W. and Zuo, M. J. Optimal reliability modeling: principles and applications. Wiley and Sons, Hoboken, NJ. 2002.

6. A. S. Moshnikov and V. S. Kolomoitcev, "Reliability Assessment of Distributed Control Systems with Network Structure," 2020 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF), 2020, pp. 1-4, doi: 10.1109/WECONF48837.2020.9131490..
7. Moshnikov, A.; Bogatyrev, V. Risk Reduction Optimization of Process Systems under Cost Constraint Applying Instrumented Safety Measures. Computers 2020, 9, 50. <https://doi.org/10.3390/computers9020050>

UDC: 517.937, 517.928.2, 519.217.2

Numerical analysis of large-scale queueing system with a small parameter

S.A.Vasilyev¹, G.O.Tsareva¹, M.A.Bouatta¹

¹Peoples' Friendship University of Russia (RUDN University),
Miklukho-Maklaya str. 6, Moscow, Russia

vasilyev-sa@rudn.ru

Abstract

In this work we study large-scale queueing systems (LSQS) with a small parameter using numerical analysis. We assume that there is a Poisson input flow of requests to LSQS with a limited intensity and there is a service discipline for any request which provides a randomly selection from any m -set servers such server that has the s -th shortest queue size. We consider Tikhonov problem for a system of differential equations with a small parameter. Solutions of Tikhonov problem are shares of the servers that have the queues lengths with not less than k . We describe the processes of rapid changes of LSQS and time scaling in this LSQS using a small parameter. We apply the adaptive numerical methods for this LSQS analysis using a piecewise-uniform grid. The results of the numerical analysis demonstrate the high efficiency of this numerical method.

Keywords: countable Markov chains, large-scale network modeling, singular perturbed systems of differential equations, small parameter, numerical analysis

1. Introduction

The research of large-scale queueing systems (LSQS) with a lager number of servers is extremely important because of the development of 5G/6G networks and Internet of Things (IoT) sets the problem of using not only analytical methods but also numerical ones [1], [4], [5]. The modern research of LSQS with complex routing discipline focused on the problems of stability analysis of infinite servers LSQS [2], [3], [9].

In this work we study LSQS with a small parameter using numerical analysis. We assume that consists of infinite number of servers with a Poisson input flow of requests of intensity $n\lambda$. Each request selects m any servers randomly and we assume that there is a next step which includes sending this request to the server that has the s -th shortest (or equivalently, the $(m - s)$ -th longest) queue size, $1 \leq s \leq m$. We

construct the Tikhonov type singular perturbed infinite order system of differential equations and we use the truncated analogue of this system in which the order of the system is very large which allows us to consider LSQS. We apply the adaptive numerical methods for this LSQS analysis using a piecewise-uniform grid. The results of the numerical analysis demonstrate the high efficiency of this numerical method.

2. Large-scale queueing systems model with a small parameter

We consider LSQS with $n \rightarrow \infty$ infinite-buffer FCFS single-services, each with its own exponentially distributed service times of mean $\bar{t} = 1/\mu$. We suppose that there is Poisson arrivals of requests with rate $\rho = n\lambda$, where $0 < \lambda < \mu$.

Assuming that we can select m servers for each request upon its arrival randomly and immediately and we can choose one server among the selected m servers that has the s -th shortest ($(m-s)$ -th longest) queue length in the choice moment. If there happen to be more than one server with the s -th shortest queue size, we select one of them randomly. The request is sent to the chosen server after this server selection procedure immediately.

Let $\mathbf{u} = \{u_k(t)\}_{k=0}^{\infty}$ be shares of the servers that have the queues lengths with not less than k , where $1 \geq u_0(t) \geq u_1(t) \geq \dots \geq u_k(t) \geq \dots$, $u_k(t) = \zeta_k/n$ ($k \in \mathbb{Z}_+$, $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$), $0 \leq \zeta_k \leq n$ are non-negative integers and $\sum_{k=1}^{\infty} u_k(t) < \infty$ for any $t \geq 0$.

As was shown, the elements of the sequences $\{u_k(t)\}_{k=0}^{\infty}$ becomes deterministic in the infinite limit $n \rightarrow \infty$ and the evolution of this large-scale system is described by solutions of an infinite system of differential equations [8]

$$\begin{cases} \dot{u}_k^{s,m}(t) = \mu(u_{k+1}^{s,m}(t) - u_k^{s,m}(t)) + \lambda(h_{s,m}(u_{k-1}^{s,m}(t)) - h_{s,m}(u_k^{s,m}(t))), \\ k \geq 1, t \geq 0, \\ u_0^{s,m}(t) = 1, u_k^{s,m}(0) = g_k \geq 0, k \geq 1, \end{cases} \quad (1)$$

where $\mathbf{g} = \{g_k\}_{k=0}^{\infty}$ ($g_0 = 1, g_k \geq g_{k+1}$) is non-increasing non-negative sequences and the function $h_{s,m}(u_k^{s,m}(t))$ has the form for $1 \leq s \leq m$ ($s, m \in \mathcal{N}$)

$$h_{s,m}(u_k^{s,m}(t)) = \sum_{l=0}^{s-1} C_m^l (1 - u_k^{s,m}(t))^l (u_k^{s,m}(t))^{m-l}.$$

When we study scale invariance in time (i.e. scales transformation of time intervals change), we can analyze the transformation properties of solutions of differential equations. Scaling transformations are similarity transformations and form a group of scale transformations.

We can investigate scaling properties of solutions for Cauchy problem of infinite system of differential equations with small parameter such form

$$\begin{cases} \epsilon^{b_k} \dot{u}_k^{s,m}(t) = \mu(u_{k+1}^{s,m}(t) - u_k^{s,m}(t)) + \lambda(h_{s,m}(u_{k-1}^{s,m}(t)) - h_{s,m}(u_k^{s,m}(t))), \\ k \geq 1, t \geq 0, \\ u_0^{s,m}(t) = 1, u_k^{s,m}(0) = g_k \geq 0, k \geq 1, g_k \geq g_{k+1}, \end{cases} \quad (2)$$

where $\epsilon > 0$ is a small parameter and $\mathbf{b} = \{b_k\}_{k=1}^{\infty}$, $(b_k \geq 0)$ is a numerical sequence of real numbers. Thus, the system (1) is the singular perturbation system and we can describe processes of rapid changes of the solutions of this system with scaling transformations this form $\bar{t}_k = \epsilon^{-b_k} t$.

This Cauchy problem (2) can be transformed into a Tikhonov problem, if we assume $b_k = 0$, $k = 1, 2, \dots, l$, $b_k > 0$, $k = l+1, l+2, \dots$ ($l \geq 2$).

3. Truncation large-scale network model and numerical analysis

We can write Tikhonov problem for the truncation system of differential equations (2)

$$\begin{cases} \dot{u}_k^{s,m}(t) = \epsilon^{-b_k} [\mu(u_{k+1}^{s,m}(t) - u_k^{s,m}(t)) + \lambda(h_{s,m}(u_{k-1}^{s,m}(t)) - h_{s,m}(u_k^{s,m}(t)))] \\ 1 \leq k \leq n, 0 \leq t \leq T, \\ u_0^{s,m}(t) = 1, u_{n+1}^{s,m}(t) = 0, u_k^{s,m}(0) = g_k \geq 0, 1 \leq k \leq n, g_k \geq g_{k+1}, \end{cases} \quad (3)$$

where $b_k = 0$, $k = 1, 2, \dots, l$, $b_k > 0$, $k = l+1, l+2, \dots, n$ ($2 \leq l \leq n$) and the parameter T is the right border of the time interval for Tikhonov problem.

We apply a piecewise-uniform grid $\bar{\Omega}_t$ ($0 = t_0 < t_1 < \dots < t_N = T$) for numerical analysis of this Tikhonov problem (3)

$$\bar{\Omega}_t = (t_i | t_i = i\tau_1; i = 0, 1, 2, \dots, K; t_i = t_K + (i-K)\tau_2; i = K+1, \dots, N),$$

$$\tau_1 = \delta/K, \tau_2 = (T-\delta)/(N-K), \delta = \bar{C}\varepsilon \ln(\varepsilon^{-1}),$$

where a parameter \bar{C} is determined by the coefficients of singularly perturbed system of differential equations. Thus, this piecewise-uniform grid $\bar{\Omega}_t$ has K small steps τ_1 and $(N-K)$ big steps τ_2 on the segment $[0, T]$.

We can consider a finite-difference approximation of the system (2) in the following form $h_i = t_i - t_{i-1}$, $u_{k+1,i}^{s,m} = u_{k+1}^{s,m}(t_i)$,

$$\begin{cases} u_{k,i+1}^{s,m} = u_{k,i}^{s,m} + h_i \epsilon^{-b_k} [\mu(u_{k+1,i}^{s,m} - u_{k,i}^{s,m}) + \lambda(h_{s,m}(u_{k-1,i}^{s,m}) - h_{s,m}(u_{k,i}^{s,m}))], \\ 1 \leq k \leq n, 0 \leq t_i \leq T, \\ u_{0,i}^{s,m} = 1, u_{n+1,i}^{s,m} = 0, u_{k,0}^{s,m} = g_k \geq 0, 1 \leq k \leq n, g_k \geq g_{k+1}. \end{cases} \quad (4)$$

We use a vector notation and apply this numerical scheme in the form

$$\begin{cases} \mathbf{u}_{i+1} = \mathbf{F}(\mathbf{u}_i, t_i) \quad i = 0, 1, \dots, N, \\ \mathbf{u}_0^{s,m} = \mathbf{g}, \end{cases} \quad (5)$$

$$\begin{aligned} \mathbf{u}_i &= (u_{1,i}^{s,m}, u_{2,i}^{s,m}, \dots, u_{n,i}^{s,m}), \mathbf{F}(\mathbf{u}_i, t_i) = (F_1(\mathbf{u}_i, t_i), F_2(\mathbf{u}_i, t_i), \dots, F_n(\mathbf{u}_i, t_i)), \\ F_k(\mathbf{u}_i, t_i) &= u_{k,i}^{s,m} + h_i \epsilon^{-b_k} [\mu (u_{k+1,i}^{s,m} - u_{k,i}^{s,m}) + \lambda (h_{s,m}(u_{k-1,i}^{s,m}) - h_{s,m}(u_{k,i}^{s,m})], \quad 1 \leq k \leq n, \\ \mathbf{u}_0^{s,m} &= (u_{1,0}^{s,m}, u_{2,0}^{s,m}, \dots, u_{n,0}^{s,m}), \mathbf{g} = (g_1, g_2, \dots, g_n). \end{aligned}$$

We use the Kutta-Merson method to calculate a numerical solution to the problem (4) and use the following formulas

$$\begin{aligned} \mathbf{q}_i^1 &= \mathbf{F}(\mathbf{u}_i, t_i), \mathbf{q}_i^2 = \mathbf{F}(\mathbf{u}_i + \frac{1}{3}\mathbf{q}_i^1, t_i + h_i/3), \mathbf{q}_i^3 = \mathbf{F}(\mathbf{u}_i + \frac{1}{6}[\mathbf{q}_i^1 + \mathbf{q}_i^2], t_i + h_i/3), \\ \mathbf{q}_i^4 &= \mathbf{F}(\mathbf{u}_i + \frac{1}{8}[\mathbf{q}_i^1 + 3\mathbf{q}_i^2], t_i + h_i/2), \mathbf{q}_i^5 = \mathbf{F}(\mathbf{u}_i + \frac{1}{2}[\mathbf{q}_i^1 - 3\mathbf{q}_i^2] + 2h_i\mathbf{q}_i^4, t_i + h_i), \\ \mathbf{u}_{i+1} &= \mathbf{u}_i + \frac{1}{6}h_i(\mathbf{q}_i^1 + 4\mathbf{q}_i^4 + \mathbf{q}_i^5), \mathbf{u}_{ci+1}^c = \mathbf{u}_i + 0.5h_i(\mathbf{q}_i^1 - 3\mathbf{q}_i^3 + 4\mathbf{q}_i^4), \end{aligned}$$

where $\mathbf{q}_i^j \in R^n \quad (i = 0, 1, \dots, N, j = \overline{1, 5})$ are vectors. $R_i = 0.2 \max_{k=\overline{1, n}} |u_{k,i+1} - u_{k,i+1}^c|$

is estimated in each point $i = \overline{1, N}$. If the value R_i is greater than the permissible error δ , then we make the step twice shorter and re-calculate the vectors \mathbf{q}_i^j .

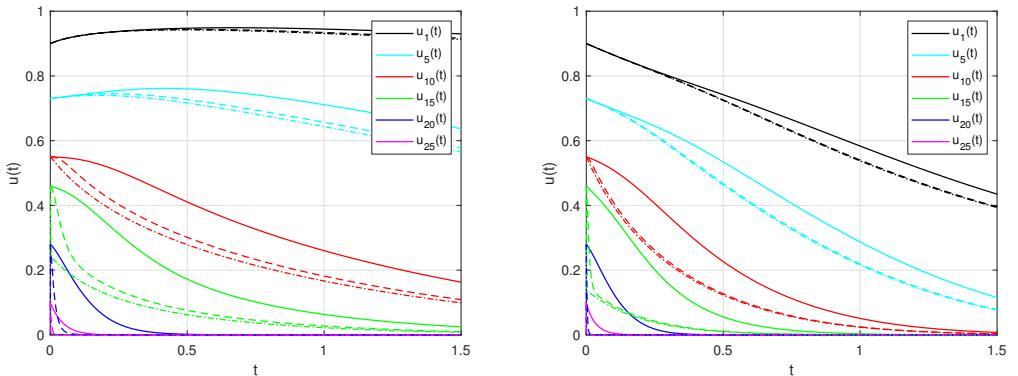


Fig. 1. Evolution analysis of $u_k^{1,2}$ (the left graph) and $u_k^{2,3}$ (the right graph) ($\epsilon = 0.1$ solid line, $\epsilon = 0.01$ long dash line, $\epsilon = 0.001$ short dash line).

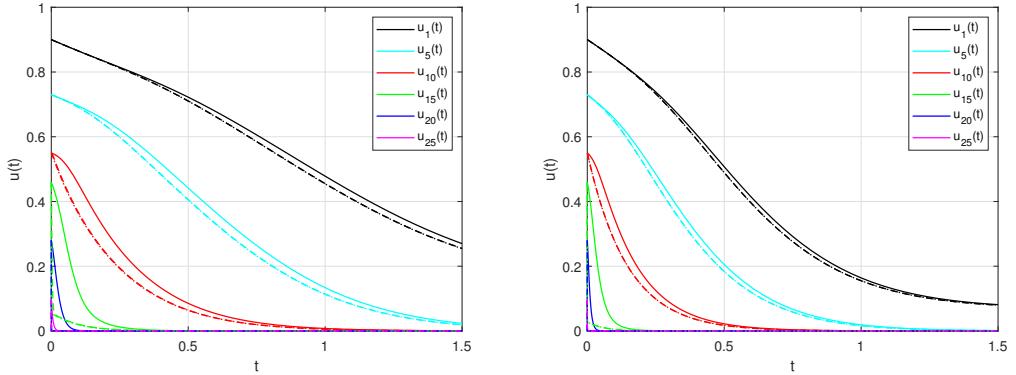


Fig. 2. Evolution analysis of $u_k^{2,4}$ (the left graph) and $u_k^{3,4}$ (the right graph) ($\epsilon = 0.1$ solid line, $\epsilon = 0.01$ long dash line, $\epsilon = 0.001$ short dash line) .

The numerical example is presented in the figure (see Fig. 1-2). We use the next parameters: $\lambda = 0.5$, $\mu = 1.1$, the dimension of the system of differential equation is $n = 25$, the number of differential equation without a small parameters $l = 9$ (i.e. $1 \leq k \leq 9$), the number of differential equation with a small parameters $n - l = 16$ (i.e. $10 \leq k \leq 25$), degrees of a small parameter are $b_k = 0, 1 \leq k \leq 9$ and $b_k = 1/k, 10 \leq k \leq 25$, the number of steps of the grid is $N = 10^4$ and the permissible error $\delta = 10^{-5}$. The value of the initial conditions are presented as follows: $g_0 = 1$, $g_k = 1 - 9k/250$, $k = \overline{1, 25}$. The parameters s, m have the following values: $s = 1, m = 2$ and $s = 2, m = 3$ on Fig. 1, $s = 2, m = 4$ and $s = 3, m = 4$ on Fig. 2.

4. Conclusion

We construct the Tikhonov type singular perturbed infinite order system of differential equations with a small parameter and use the truncated analogue of this system. The truncation procedure allow us to apply adaptive numerical methods for solution analysis of the Tikhonov problem using a piecewise-uniform grid. The results of the numerical analysis demonstrate the high efficiency of this numerical method. We consider LSQS with different parameters of the service disciplines. It is shown that there are time-scale effects in the behavior of solutions. It is demonstrate the appearance of boundary layers solution of the equation with a small parameter. The decrease in the value of the solutions over time is clearly visible on all graphs. We can see that LSQS has a more efficient servicing of input flow of requests, if we increase the parameter m . We suppose that it is possible to formulate an optimal control problem for such LSQS and successfully solve it. Our ability to solve such

problems is very important because it will save technical and financial resources for 5G/6G networks implementation. This paper has been supported by the RUDN University Strategic Academic Leadership Program and it has been funded by RFBR, project number 18-07-00567.

REFERENCES

1. van der Boor, M., Borst, S., van Leeuwaarden, J. 2021. "Optimal hyper-scalable load balancing with a strict queue limit". *Performance Evaluation*, Vol. 149–150, p. 102217. <https://doi.org/10.1016/j.peva.2021.102217>.
2. Fiems, D., Mandjes, M., Patch, B. 2018. "Networks of infinite-server queues with multiplicative transitions". *Performance Evaluation*, Vol. 123–124, p. 35–49. <https://doi.org/10.1016/j.peva.2018.03.003>.
3. Jansen, H.M., Mandjes,M., De Turck, K., Wittevrongel,S. 2019. " Diffusion limits for networks of Markov-modulated infinite-server queues". *Performance Evaluation*, Vol. 135, p. 102039. <https://doi.org/10.1016/j.peva.2019.102039>.
4. van Kreveld, L.R., Boxma,O.J., Dorsman, J.L., Mandjes,M.R.H. 2021. " Scaling limits for closed product-form queueing networks". *Performance Evaluation*, Vol. 151, p. 102220. <https://doi.org/10.1016/j.peva.2021.102220>
5. Stefan Rank, S., Hans-Peter Schwefel, H.-P. 2006 "Transient analysis of RED queues: A quantitative analysis of buffer-occupancy fluctuations and relevant time-scales". *Performance Evaluation*, Vol. 63, Issue 8, pp. 725–742. <https://doi.org/10.1016/j.peva.2005.08.001>
6. Vvedenskaya, N.D., Dobrushin, R.L., Kharpelevich, F.I. 1996. "Queueing system with a choice of the lesser of two queues — the asymptotic approach". *Probl. inform.*, Vol. 32, Issue 1, pp.15–27.
7. Vvedenskaya, N.D., Suhov, Yu.M. 1997. "Dobrushin's Mean-Field Approximation for a Queue with Dynamic Routing". *Markov Processes and Related Fields*, Issue 3, pp. 493–526.
8. Vvedenskaya, N.D. 1998. "A large queueing system with message transmission along several routes". *Problemy Peredachi Informatsii*, Vol. 34, no. 2, pp. 98–108.
9. Yajima, M., Phung-Duc, T. 2019. "A central limit theorem for a Markov-modulated infinite-server queue with batch Poisson arrivals and binomial catastrophes". *Performance Evaluation*, Vol. 129, pp. 2–14. <https://doi.org/10.1016/j.peva.2018.10.002>.

UDC: 519.248

On the reliability estimation of the FBM multi-phase degradation system

O.V. Lukashenko^{1,2}

¹Institute of Applied Mathematical Research of the Karelian Research Centre of RAS,
Petrozavodsk, Russia

²Petrozavodsk State University, Petrozavodsk, Russia

lukashenko@krc.karelia.ru

Abstract

The estimation of the reliability is an important and hard problem, arising in the performance analysis of degradation systems. The required performance measure is usually not analytically available. Hence, one has to rely on simulation methods. The variance reduction technique is developed to estimate the reliability of the multi-phase degradation model based on fractional Brownian motion (FBM). Numerical experiments are conducted to evaluate the performance of the proposed estimator.

Keywords: Reliability, Degradation process, Fractional Brownian motion, Conditional Monte Carlo, Bridge process

1. Introduction

Degradation is the main cause of infrastructure failure in technical and engineering systems. Investigation of the degradation model performance characteristics is a very important problem in reliability analysis aimed at the prediction of fault tolerance. Thus, the development and evaluation of the models describing the degradation process is an actual research area. In this framework the Wiener process is one of the most popular degradation models (see [1] and references therein) thanks to its analytical tractability. There are some extensions on the general Gaussian process [2].

The standard models assume the fixed failure threshold which can be not realistic in practice, since deterioration system operates in a random environment and the thresholds and other parameters can change dynamically. A multi-phase degradation model based on the Wiener process with changing drift and threshold parameters was developed in [3]. In this paper we focus on the general case of fractional Brownian motion with correlated increments for which analytical results are not available and

the required performance measure is estimated via simulation. Note that FBM was considered as a degradation model in [4, 5] but they didn't consider the multi-phase case.

By definition failure occurs when the degradation process reaches the predefined thresholds. For systems with high reliability, the failures may not occur during a short time under normal conditions. Since the event of interest becomes rare, a special case of Conditional Monte Carlo, based on the bridge process is proposed.

The rest of this paper is organized as follows. Section 2 describes the degradation model in detail. In Section 3, the conditional Monte Carlo method based on the Bridge process is proposed to estimate the reliability of the considered system. A simulation study is carried out in Section 4 to demonstrate the performance of the proposed estimator. Finally, some conclusions are stated in Section 5.

2. Model description and performance measures

Let consider the so-called multi-phase degradation model with a deterministic sequence of the change points $\tau_1 < \dots < \tau_n$. Each time period (τ_{i-1}, τ_i) , $i = 1, \dots, n$ is characterized by the failure threshold level D_i and the mean degradation rate m_i . Assume that the degradation dynamic of the considered deterioration system satisfies the following model

$$A(t) = \Lambda(t) + X(t), \quad (1)$$

where random fluctuations are described in terms of the fractional Brownian motion (FBM) $\{X(t), t \geq 0\}$ with a Hurst parameter $H \in (0, 1)$ and covariance function

$$\Gamma(t, s) := \mathbb{E}[X(t)X(s)] = \frac{1}{2} (t^{2H} + s^{2H} - |t - s|^{2H}).$$

The deterministic drift $\Lambda(t)$ is a piece-wise constant function, namely

$$\Lambda(t) = \sum_{i=1}^n m_i \cdot t \cdot I(\tau_{i-1} < t < \tau_i),$$

where I denotes the indicator function.

Let denote the failure threshold set $\mathcal{D} = (D_1, \dots, D_n)$. Then, the lifetime of the deterioration system is

$$T_{\mathcal{D}} := \min\{t : A(t) \geq D(t)\}, \quad (2)$$

i.e., the first time the process $\{X(t)\}$ hits the threshold piece-wise constant curve

$$D(t) = \sum_{i=1}^n D_i I(\tau_{i-1} < t < \tau_i).$$

The reliability of the system is defined as the tail distribution of the lifetime $T_{\mathcal{D}}$:

$$R(u | \mathcal{D}) := \mathbb{P}(T_{\mathcal{D}} \geq u). \quad (3)$$

The typical sample path of the degradation process of a deterioration system under changing thresholds is shown in Fig. 1.

Analytical expression of system reliability is known for the Wiener process ($H = 0.5$) model with two failure thresholds [3]. For the multiple failure threshold a recursive formula is developed. In general FBM case the analytical solution is not available in closed form. Note that in single-phase case when $D_1 = D_2 = \dots = D_n$, $m_1 = m_2 = \dots = m_n$ (which correspond to the hitting time distribution evaluation) there are some bounds and asymptotic results [6, 7]

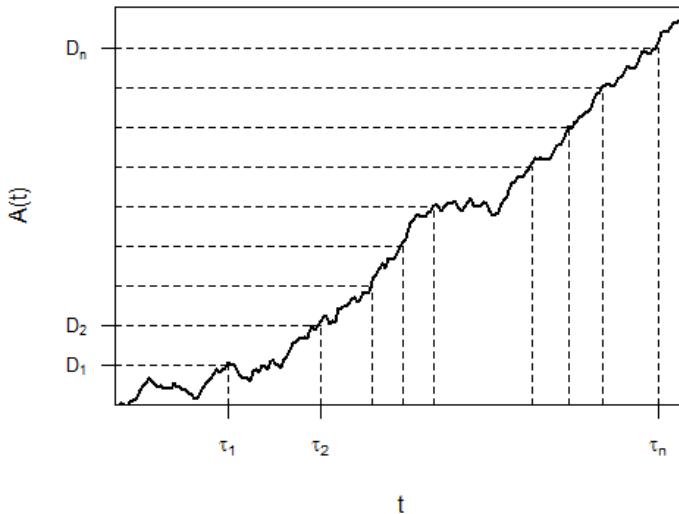


Fig. 1. The typical sample path of the degradation process.

3. Monte Carlo estimation

In this paper, our task is to estimate the system reliability for the degradation system with changing failure thresholds and mean rates via Monte Carlo (MC) simulation.

The standard MC approach is based on the estimation of the mean indicator of the event of interest:

$$R(u | \mathcal{D}) = \mathbb{E}I(T_{\mathcal{D}} \geq u), \quad (4)$$

The main problem is that for large values u the target probability is extremely small, hence a standard MC estimate fails to evaluate the quantity of interest with given accuracy (usually expressed in terms of the relative error). Thus, a special rare-event simulation technique is required. In this paper, we apply a special variant of the Conditional MC (CMC) method [8, 9]. Under this approach, a conditional expectation of the target quantity with respect to some auxiliary random variable is estimated, provided this expectation is available in closed form. In general, it is not easy to select an appropriate auxiliary random variable, however in the current setting it can be expressed in terms of the so-called bridge process [10, 11]

$$Y(t) = X(t) - \psi(t)X(s), \quad (5)$$

where s is a some prefixed instant, ψ is expressed in terms of the covariance function of the process X :

$$\psi(t) := \frac{\Gamma(t, s)}{\Gamma(s, s)}.$$

The target probability can be rewritten in terms of the corresponding bridge process as follows:

$$\begin{aligned} R(u | \mathcal{D}) &= \mathbb{P}(T_{\mathcal{D}} \geq T) \\ &= \mathbb{P}(\forall t \in [0, u] : \Lambda(t) + X(t) \leq D(t)) \\ &= \mathbb{P}\left(\forall t \in [0, u] X(s) \leq \frac{D(t) - Y(t) - \Lambda(t)}{\psi(t)}\right) \\ &= \mathbb{P}(X(s) \leq \bar{Y}), \end{aligned}$$

where

$$\bar{Y} := \inf_{t \in [0, u]} \frac{D(t) - Y(t) - \Lambda(t)}{\psi(t)}, \quad (6)$$

and, due to the independence between \bar{Y} and $X(s)$ and the properties of the conditional expectation:

$$R(u | \mathcal{D}) = \mathbb{P}(X(s) \leq \bar{Y}) = \mathbb{E}\left[\Psi\left(\frac{\bar{Y}}{\sqrt{\Gamma(s, s)}}\right)\right],$$

where Ψ denotes the distribution function of a standard normal variable.

Hence, given N independent samples $\{\bar{Y}^{(n)}, n = 1, \dots, N\}$ of \bar{Y} , the BMC estimator of $R(u | \mathcal{D})$ is

$$\hat{R}_N^{\text{BMC}} := \frac{1}{N} \sum_{n=1}^N \Psi \left(\frac{\bar{Y}^{(n)}}{\sqrt{\Gamma(s, s)}} \right). \quad (7)$$

4. Simulation Results

In this section, we provide a preliminary simulation analysis of the accuracy of the BMC estimator for the two-phase degradation with the mean degradation rates $m_1 = 1, m_2 = 1.2$ and the failure thresholds $D_1 = 20, D_2 = 22$. $N = 10000$ replications of the FBM with Hurst parameter $H = 0.7$ were generated. The conditioning point $s = \lfloor u/2 \rfloor$. To verify the effectiveness of the proposed estimator, we considered the dependence of the relative error (defined as coefficient of variation of the estimator) on the rarity parameter u in comparison with the standard MC estimator. The numerical results are presented in the Table ??.

Table 1. Performance of the estimators for the FBM with $H = 0.7$.

u	\hat{R}^{MC}	\hat{R}^{BMC}	$\text{RE}(\hat{R}^{\text{MC}})$	$\text{RE}(\hat{R}^{\text{BMC}})$
50	6.1e-03	6.16e-03	1.27e-01	2.28e-03
70	6e-04	6.17e-04	4.08e-01	3.29e-03
90	2e-04	8.56e-05	7.07e-01	4.03e-03
110	–	1.48e-05	–	4.62e-03
130	–	3.01e-06	–	5.23e-03
150	–	6.98e-07	–	5.64e-03
170	–	1.79e-07	–	5.99e-03
190	–	4.97e-08	–	6.31e-03
210	–	1.46e-08	–	6.69e-03
230	–	4.57e-09	–	7.07e-03

5. Conclusion

In this paper, we have proposed the BMC estimation of the reliability of the degradation system with multiple phases with different mean degradation rates and failure thresholds. In more detail, we have derived the expression of the estimator and provided some preliminary simulation results, highlighting how a limited number of replications ($N = 10^4$) permits to estimate probabilities of the order of 10^{-9} with comparatively moderate value of the relative error. One of the possible research direction is studying the sensitivity of the reliability estimation to the values of the technical parameters.

Acknowledgment

The study was supported by the Russian Science Foundation, project 21-71-10135.

REFERENCES

1. E. S. Chetvertakova, E. V. Chimitova, The Wiener degradation model in reliability analysis, in: 2016 11th International Forum on Strategic Technology (IFOST), 2016, pp. 488–490.
2. Z. Wang, Q. Wu, X. Zhang, X. Wen, Y. Zhang, C. Liu, H. Fu, A generalized degradation model based on gaussian process, *Microelectronics Reliability* 85 (2018) 207–214. doi:10.1016/j.microrel.2018.05.001.
3. H. Gao, L. Cui, D. Kong, Reliability analysis for a Wiener degradation process model under changing failure thresholds, *Reliability Engineering & System Safety* 171 (2018) 1–8. doi:10.1016/j.ress.2017.11.006.
4. H. Zhang, D. Zhou, M. Chen, X. Xi, Predicting remaining useful life based on a generalized degradation with fractional brownian motion, *Mechanical Systems and Signal Processing* 115 (2019) 736–752. doi:10.1016/j.ymssp.2018.06.029.
5. H. Zhang, M. Chen, J. Shang, C. Yang, Y. Sun, Stochastic process-based degradation modeling and rul prediction: from brownian motion to fractional brownian motion, *Science China Information Sciences* 64 (7) (2021) 171201. doi:10.1007/s11432-020-3134-8.
6. Z. Michna, On tail probabilities and first passage times for fractional brownian motion, *Mathematical Methods of Operations Research* 49 (1999) 335–354.
7. M. Caglar, C. Vardar, Distribution of maximum loss of fractional brownian motion with drift, *Statistics and Probability Letters* 83 (2013) 2729–2734.
8. S. M. Ross, *Simulation*, Elsevier, 2006.
9. D. P. Kroese, T. Taimre, Z. I. Botev, *Handbook of Monte Carlo Methods*, John Wiley & Sons, 2011.
10. S. Giordano, M. Gubinelli, M. Pagano, Bridge Monte-Carlo: a novel approach to rare events of Gaussian processes, in: Proc. of the 5th St.Petersburg Workshop on Simulation, St. Petersburg, Russia, 2005, pp. 281–286.
11. S. Giordano, M. Gubinelli, M. Pagano, Rare events of Gaussian processes: a performance comparison between Bridge Monte-Carlo and Importance Sampling, in: Next Generation Teletraffic and Wired/Wireless Advanced Networking, St. Petersburg, Russia, 2007, pp. 269–280.

UDC: 519.218, 519.873

Stability of Some Applied Probability Models

E.V.Bulinskaya¹

¹Lomonosov Moscow State University, Leninskie Gory 1, 119991 Moscow, Russia
ebulinsk@mech.math.msu.su

Abstract

In order to study a real process or system researcher has to choose an appropriate mathematical model. Usually one considers input-output models dealing with such probability theory applications as communication, reliability, queueing, inventory, finance, insurance and others. Modern period in actuarial sciences is characterized by consideration of complex systems, interplay of cost and reliability approaches and wide use of computers, therefore it is desirable to treat discrete-time models. They can give approximation to corresponding continuous-time ones. Moreover, in some cases they provide a better description of a real-life situation. So, we consider two discrete-time models and investigate their optimal performance and stability.

Keywords: Stability, Reliability, Applied probability models

1. Introduction

The models used in applications of probability theory are for the most part of input-output type. They are described by specifying the planning horizon, input and output processes and objective function. According to the choice of objective function there exist two main approaches (cost and reliability ones), see, e.g., [1]. We consider two discrete-time models. The first one, illustrating the cost approach, describes the performance of insurance company using proportional reinsurance and bank loans. The second model treats functioning of the company employing investment in non-risky asset for a fixed period. An algorithm for calculation of ultimate ruin probability is obtained. Thus, reliability approach is used in this case.

Discrete-time models became popular during the last decades. On the one hand, they give an approximation to corresponding continuous-time models (see, e.g., [2]), on the other, sometimes provide a better description of a real-life situation. One of the first papers concerning discrete models in ruin theory [3] appeared in 1988. The

The publication has been prepared with the support of RFBR according to the research project No.20-01-00487.

paper [4] gives a review of discrete-time models considered until 2009. A review of recent results on discrete-time models is supplied, as well, in Section 5 of [5]. It is important to mention that for obtaining the optimal behavior of a risk model one can use such decisions as investment, bank loan or reinsurance, see, e.g., [6]-[8].

Due to lack of space we omit below some proofs.

2. Discrete-time model with reinsurance and bank loans

2.1. Model description. Suppose that the claims arriving to insurance company are described by a sequence of independent identically distributed (i.i.d.) non-negative random variables (r.v.'s) $\{X_i, i \geq 1\}$. Here X_i is the claim amount during the i -th period (year, month or day). Let $F(x)$ be its distribution function (d.f.) having density $\varphi(x)$ and finite expectation. The company uses proportional reinsurance with quota α , that is, pays only αX if the claim amount is X , and bank loans. If a loan is taken at the beginning of period (before the claim arrival) the rate is k whereas the loan after the claim arrival is taken at the rate r with $r > k$. Our aim is to choose the loans in such a way that the additional payments entailed by loans are minimized. Denote by M the premium acquired by direct insurer (after reinsurance) during each period. If x is the initial capital then $f_1(x)$, the minimal expected additional costs during one period, are given by

$$f_1(x) = \min_{y \geq x} [k(y - x) + rE(\alpha X - y)^+], \quad (1)$$

here y is the company capital after the loan and $(\alpha X - y)^+ = \max(\alpha X - y, 0)$. Clearly, (1) can be rewritten in the form:

$$f_1(x) = -kx + \min_{y \geq x} G_1(y), \quad \text{with} \quad G_1(y) = ky + r\alpha \int_{y/\alpha}^{\infty} \bar{F}(s) ds. \quad (2)$$

Now let $f_n(x)$ be the minimal expected costs during n periods and β the discount factor for future costs, $0 < \beta < 1$. Then using the dynamic programming (see, e.g., [9]), one easily obtains the following relation:

$$f_n(x) = -kx + \min_{y \geq x} G_n(y), \quad \text{with} \quad G_n(y) = G_1(y) + \beta E f_{n-1}(y + M - \alpha X). \quad (3)$$

2.2. Optimization problem.

Theorem 1. There exists an increasing sequence of critical levels $\{y_n\}_{n \geq 1}$ such that

$$f_n(x) = -kx + \begin{cases} G_n(y_n), & \text{if } x \leq y_n, \\ G_n(x), & \text{if } x > y_n. \end{cases} \quad (4)$$

Let \bar{y} satisfy $H(\bar{y}) = 0$ where $H(y) = G'_1(y) - k\beta$, then $y_n \leq \bar{y}$.

Proof. Consider $G'_1(y) = 0$ where $G'_1(y) = k - r\bar{F}(y/\alpha)$. Since $r > k$, it follows immediately that $y_1 = \alpha F^{-1}(1-k/r)$ exists and is the unique solution of the equation under consideration.

Further results are obtained by induction. Since $f_1(x)$ is given by (4),

$$f'_1(x) = -k + \begin{cases} 0, & \text{if } x \leq y_1, \\ G'_1(x), & \text{if } x > y_1, \end{cases} = \begin{cases} -k, & \text{if } x \leq y_1, \\ -r\bar{F}(y/\alpha), & \text{if } x > y_1. \end{cases} \quad (5)$$

Hence, it is clear that $f'_1(x) < 0$ for all x . Moreover, on the one hand,

$$G'_2(y) = G'_1(y) + \beta \int_0^\infty f'_1(y + M - \alpha s)\varphi(s) ds \leq G'_1(y),$$

on the other hand, $G'_2(y)$ has the form

$$H(y) + \beta \int_0^{\frac{y+M-y_1}{\alpha}} G'_1(y + M - \alpha s) ds, \text{ with } H(y) = G'_1(y) - \beta k.$$

That means $y_1 < y_2 < \bar{y}$. Furthermore, $f'_2(x) < 0$ for all x . Thus, the base of induction is established. Assuming that (4) is true for the number of periods less or equal to n we prove its validity for $n + 1$ and deduce that $G'_{n+1}(x) < G'_n(x)$, so $y_n < y_{n+1}$. On the other hand,

$$G'_{n+1}(y) = H(y) + \beta \int_0^{\frac{y+M-y_n}{\alpha}} G'_n(y + M - \alpha s) ds,$$

hence $G'_{n+1}(y) > H(y)$. This entails the needed relation $y_{n+1} < \bar{y}$. ■

Corollary 1. There exists $\hat{y} = \lim_{n \rightarrow \infty} y_n$.

Remark 1. It is interesting that $\hat{y} = \bar{y}$ only for $M = 0$ whereas $\hat{y} < \bar{y}$ for $M > 0$.

2.3. Model stability. Now we turn to sensitivity analysis and prove that the model under consideration is stable with respect to small perturbations of the underlying distribution. For this purpose we introduce two variants of the model (with claims X_i and Y_i). In the first one all functions has subscript X and in the second one Y .

For random variables X and Y defined on some probability space and possessing finite expectations it is possible to define their distance on the base of Kantorovich metric in the following way

$$\varkappa(X, Y) = \int_{-\infty}^{\infty} |F_X(t) - F_Y(t)| dt$$

where F_X and F_Y are the distribution functions of X and Y respectively.

The distance between the cost functions is measured in terms of Kolmogorov uniform metric. Thus, we study

$$\Delta_n = \sup_x |f_{n,X}(x) - f_{n,Y}(x)|.$$

To this end we need the following

Lemma 1. Let functions $g_i(y)$, $i = 1, 2$, be such that $|g_1(y) - g_2(y)| < \delta$ for some $\delta > 0$ and any y , then $\sup_x |\inf_{y \geq x} g_1(y) - \inf_{y \geq x} g_2(y)| < \delta$.

Now we are able to estimate Δ_1 .

Lemma 2. Assume $\varkappa(X, Y) = \rho$, then $\Delta_1 \leq \alpha r \rho$.

Proof. According to Lemma 1 we need to estimate $|G_{1,X}(y) - G_{1,Y}(y)|$ for any y . The definition of these functions gives $G_{1,X}(y) - G_{1,Y}(y) = r[E(\alpha X - y)^+ - E(\alpha Y - y)^+] = r\alpha \int_{y/\alpha}^{\infty} (\bar{F}_X(t) - \bar{F}_Y(t)) dt$. This leads immediately to the desired estimate. ■

Next, we prove the main result demonstrating the model's stability.

Theorem 2. If $\varkappa(X, Y) = \rho$, then $\Delta_n \leq D_n \rho$ where $D_n = \alpha \left(\frac{r(1-\beta^n)}{1-\beta} + \frac{k(\beta-\beta^n)}{1-\beta} \right)$.

Proof. As in Lemma 2 we estimate $|G_{n,X}(y) - G_{n,Y}(y)|$ for any y . Due to definition (3) we have

$$\begin{aligned} |G_{n,X}(y) - G_{n,Y}(y)| &\leq |G_{1,X}(y) - G_{1,Y}(y)| \\ &+ \beta \left| \int_0^\infty f_{n-1,X}(y + M - \alpha s) \varphi_X(s) ds - \int_0^\infty f_{n-1,Y}(y + M - \alpha s) \varphi_Y(s) ds \right| \end{aligned}$$

Obviously, the first term on the right-hand side of inequality is less than $\alpha r \rho$. To estimate the second term we rewrite it adding and subtracting an integral with different indices of f_{n-1} and φ . Then integrating by parts and using $\max_y |f'_{n-1,Y}(y)| \leq k$ for all n , we get

$$\Delta_n \leq \alpha(r + k\beta)\rho + \beta\Delta_{n-1}.$$

Solving this recurrent relation leads to the necessary form of D_n . ■

Corollary 2. $\Delta_n \leq \frac{r+k\beta}{1-\beta} \alpha \rho$ for any n .

In other words, we established stability of the model with respect to small perturbations of claim distribution.

3. Model with investment in a non-risky asset

3.1. Model description. We consider the following generalization of the model introduced in [10] and further treated in [11]. For certainty, we proceed in terms of insurance company performance. During the i th period the company gets a fixed premium amount c and pays a random indemnity X_i , $i = 1, 2, \dots$. It is supposed that $\{X_i\}_{i \geq 1}$ is a sequence of non-negative independent r.v.'s having Erlang(2, λ_i) distributions. Let the initial capital $S_0 = x > 0$ be fixed. It is possible to place a quota δ of this amount in a bank for m periods, the interest rate being β per period. Thus, the surplus at the end of the first period has the form $S_1 = (1 - \delta)S_0 + c - X_1$. The same procedure is repeated each period. Putting $u_m = (1 + \beta)^m$ we get the following recurrent formula

$$S_n = \min[(1 - \delta)S_{n-1}, S_{n-1}] + c + u_m \delta S_{n-(m+1)}^+ - X_n. \quad (6)$$

Here and further on it is assumed that $S_k = 0$ for $k < 0$. The expression (6) is useful if it is permitted to delay the company insolvency (Parisian ruin, see, e.g., [12]). If we use the classical notion of ruin, then the ruin time τ is defined as follows

$$\tau = \inf\{n > 0 : S_n \leq 0\}. \quad (7)$$

To calculate the ultimate ruin probability

$$P(\tau < \infty) = \sum_{n=1}^{\infty} P(S_1 > 0, \dots, S_{n-1} > 0, S_n \leq 0)$$

one can use instead of (6) the relation

$$S_n = (1 - \delta)S_{n-1} + c - X_n + u_m \delta S_{n-(m+1)}, \quad (8)$$

treated in [10] for $m = 1$.

There is no need to find the explicit form of S_n (although it is possible to obtain it). For further investigation we need only the following result.

Lemma 3. Put $S_0 = x$ and

$$S_n = f_n - \sum_{i=1}^n g_{n,i} X_i, \quad n \geq 1, \quad (9)$$

then $f_0 = x$, whereas

$$f_n = (1 - \delta)f_{n-1} + c, \quad n = \overline{1, m}, \quad f_n = (1 - \delta)f_{n-1} + u_m \delta f_{n-(m+1)} + c, \quad n > m, \quad (10)$$

and

$$\begin{aligned} g_{n,i} &= (1 - \delta)g_{n-1,i} + u_m \delta g_{n-(m+1),i}, \quad i = \overline{1, n-(m+1)}, \\ g_{n,i} &= (1 - \delta)g_{n-1,i}, \quad i = \overline{n-m, n-1}, \quad g_{n,n} = 1. \end{aligned} \quad (11)$$

Proof. Obviously, combination of (8) and (9) provides the desired recursive relations (10) and (11). \blacksquare

3.2. Algorithm for ruin probability calculation. Our aim is investigation of ultimate and finite-time ruin probabilities. To this end, we reformulate Lemma 3 as follows. The company surplus S_n , $n \geq 1$, has the form $S_n = f_n - Y_n$ with f_n defined by (10) and vector $\mathbf{Y}_n = (Y_1, \dots, Y_n)$ given by $\mathbf{Y}_n = \mathbf{G}_n \cdot \mathbf{X}_n$. Here vector $\mathbf{X}_n = (X_1, \dots, X_n)$ and matrix $\mathbf{G}_n = (g_{k,i})_{k,i=1,\dots,n}$, where $g_{k,i} = 0$ for $i > k$, the others being specified by (11).

Theorem 3. The ultimate ruin probability is represented by

$$\sum_{n=1}^{\infty} \int_0^{f_1} \cdots \int_0^{f_{n-1}} \int_{f_n}^{+\infty} \prod_{k=1}^n p_{X_k}(v_k(y_1, \dots, y_n)) dy_1 \dots dy_n,$$

where f_n , $n \geq 1$, and $g_{n,i}$, $n \geq 1$, $i = \overline{1, n}$, are given by (10) and (11), respectively. The function $v_k(y_1, \dots, y_n)$, $k = \overline{1, n}$, is the k th component of vector $\mathbf{G}_n^{-1} \mathbf{y}_n$.

Proof. The ruin time is defined by (7). So, we calculate the probability of ruin at the n th step obtaining the distribution of the ruin time. To this end, we use Lemma 3. Hence, the probability under consideration $P(\tau = n)$ is given by $P(U_n)$ with

$$U_n = \left\{ g_{1,1}X_1 < f_1, \dots, \sum_{i=1}^{n-1} g_{n-1,i}X_i < f_{n-1}, \sum_{i=1}^n g_{n,i}X_i \geq f_n \right\}.$$

Performing the change of variables $Y_k = \sum_{i=1}^k g_{k,i}X_i$, $1 \leq k \leq n$, we obtain a one-to-one correspondence between $\mathbf{Y}_n = (Y_i)_{i=1}^n$ and $\mathbf{X}_n = (X_i)_{i=1}^n$. Since $\det \mathbf{G}_n = 1$, we get the form of ruin probability. \blacksquare

4. Conclusion and further research directions

We have studied two discrete-time models. In the first one we used the cost approach and obtained the loan strategy minimizing expected n -step costs. The model's stability with respect to small distribution perturbations is established. Another interesting result which will be presented during the talk pertains to the model with reinsurance and investment in risky and non-risky assets.

In the second model we used the reliability approach, namely, studied the ruin probability (ultimate and finite-time) under assumption that claims distributions have Erlang2 type instead of being exponential. The sensitivity analysis is also performed. It is planned to carry out the investigation of dividends problems.

REFERENCES

1. Bulinskaya E.-V. Cost approach versus reliability // Proceedings of International Conference DCCN-2017. Technosphere. Moscow. P. 382–389.
2. Dickson D.-C.-M., Waters H.-R. Some optimal dividends problems // Astin Bull. 2004. V. 34. P. 49–74.
3. Gerber H.-U. Mathematical fun with compound binomial process // Astin Bull. 1988. V. 18(2). P. 161–168.
4. Li S., Lu Y. and Garrido J. A review of discrete-time risk models // Revista de la Real Academia de Ciencias Naturales. Serie A, Matemáticas. 2009. V. 103. P. 321–337.
5. Bulinskaya E. New research directions in modern actuarial sciences // Springer Proceedings in Mathematics and Statistics. 2017. V. 208. P. 349–408.
6. Paulsen J. Ruin models with investment income // arXiv:0806.4125v1(math.PR) 25 Jun 2008.
7. Bulinskaya E., Gusak J., Muromskaya A. Discrete-time insurance model with capital injections and reinsurance // Methodology and Computing in Applied Probability. 2015. V. 17. P. 899–914.
8. Bulinskaya E. Asymptotic analysis of insurance models with bank loans // New perspectives on stochastic modeling and data analysis. Bozeman J.-R., Girardin V., Skiadas Ch. (eds.) ISAST, Athens, Greece, 2014. P. 255–270.
9. Bellman R. Dynamic programming. Princeton University Press, Princeton, New Jersey, 1957.
10. Bulinskaya E.V., Kolesnik A.D. Reliability of a discrete-time system with investment //DCCN 2018, Springer Book : Distributed Computer and Communication Networks, Chapter No: 31, 2018. P. 365–376.
11. Bulinskaya E., Shigida B. Discrete-Time Model of Company Capital Dynamics with Investment of a Certain Part of Surplus in a Non-Risky Asset for a Fixed Period // Methodology and Computing in Applied Probability. 2021. V. 23. P. 103–121.
12. Czarna I., Palmovsky Z., Świątek P. Discrete time ruin probability with Parisian delay // arXiv:1403.7761v2[math.PR], 14 Jun 2017.

UDC: 621.396

Analytical model of data transmission through NarrowBand-IoT technology

P. Keyela¹, I.S. Yartseva¹, Yu.V. Gaidamaka^{1,2}

¹Applied Probability and Informatics Department, Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, 117198, Moscow, Russia

²Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44/2 Vavilova St, 119333, Moscow, Russia

keyela17@yahoo.fr, 1032193012@rudn.ru, gaydamaka-yuv@rudn.ru

Abstract

The Internet of Things is a new stage in the development of the Internet, greatly expanding the possibilities for collecting, analyzing and sharing data. This paper explores the new innovative technology of the Internet of Things, the narrowband-IoT technology, which is rising as a solution to the multiple challenges of current mobile technologies. Specifically we consider a coverage area of a base station to which sensors in that area send data through NB-IoT technology. We build a model of sensors and base station interactions as a queueing system and analytically assess some performance metrics such as message loss probability, message waiting delay, sensor's energy consumption and sensor's lifetime.

Keywords: NB-IoT technology, IoT devices, Markov Chains, message loss, waiting delay, energy consumption, sensor's lifetime

1. Introduction

With the development of the Internet of Things (IoT), the number of connections to mobile network operators is increasing every year, and current mobile technologies face difficulties in handling that increase, due to insufficient coverage, the high cost of end terminals and the short lifetime of their batteries [1, 2]. Narrowband Internet of Things (NB-IoT) technology is an innovative solution to these challenges. It is a wireless narrowband variety of low power wide area networks (LPWANs) that is primarily intended for machine-to-machine (M2M) applications. The NB-IoT standard will open up a wide range of new opportunities for companies specializing in the provision of telecommunication services [3–5]. NB-IoT technology will occupy its low-speed niche in a class of solutions where uninterrupted data transmission and

low power consumption are a priority. In this paper we are building a mathematical model for the rapid transmission of data through NB-IoT technology and studying some quality indicators.

2. System model

Let's consider the scenario of rapid transmission of messages through the NB-IoT technology [6, 7]. We study a model in which the coverage area of the base station (BS) and the NB-IoT sensors work in an autonomous mode. The model for the analysis of performance indicators of the NB-IoT technology is constructed in the form of a discrete Markov Chain (MC). A sensor generates the message at a random time instant and notifies the BS of the need for data transmission. The network scheduler accepts the request and sends it to the queue to wait for the instant when the K channels are released from data transmission. The presence of buffer with the capacity of receiving up to L messages at the base station makes it possible to control the transmission delays and the sensor's power consumption. After the request is accepted in the buffer, the sensor switches to listening mode all the time waiting before data transmission begins. If upon arriving a message there is no place in the queue, that message is considered lost. As soon as it becomes possible to transmit the data, the BS allocates a resource block (RB) to the sensor, which starts transmitting its data with a given power level. Note that similar task arise when studying efficiency of data exchange between drones in a swarm.

3. Mathematical modeling of the system

We build our model as a discrete-time queueing system with batch arrivals and batch service with time slot length Δ . The incoming flow is Poissonian, hence the number of messages arriving in a time slot is a random variable distributed according to Poisson's law with the parameter $\lambda\Delta$. We consider that the states of the system change immediately after the end of each time slot as a result of the departure of the messages to be served [8, 9]. The probability that k messages of Poissonian incoming flow will arrive during time interval of length Δ is given by

$$A_k = \frac{(\lambda\Delta)^k}{k!} e^{-\lambda\Delta}, \quad k = 0, 1, 2, \dots \quad (1)$$

Let's introduce the following notations [10]: $Q(n)$ – the number of messages in the queue in the beginning of the n -th time slot, $A(n)$ – the number of messages received in the n -th time slot, and $S(n)$ – the number of messages served in the n -th time slot.

Recall L being the buffer capacity and K is the number servers, then $\tilde{Q}(n) = \min(Q(n) + A(n), L)$, $n \geq 1$ - the number of messages in the queue in the end of the

n -th time slot, $\tilde{Q}(n) = 0$, $S(n) = \min(\tilde{Q}(n), K)$ and the value of $Q(n)$ can be found by

$$Q(n) = \tilde{Q}(n-1) - S(n-1), \quad n = 1, 2, \dots \quad (2)$$

The state space \mathcal{Q} of the Markov Chain $Q(n)$ is $\mathcal{Q} = \{0, 1, 2, \dots, L-K\}$ and the matrix of transition probabilities $\mathbf{P} = [p_{ij}]_{i,j \in \mathcal{Q}}$ has the following form:

$$1. \quad 0 \leq i \leq K$$

$$p_{ij} = \begin{cases} \sum_{k=0}^{K-i} A_k, & j = 0; \\ A_{K+j-i}, & 0 < j < L-K; \\ \sum_{k=L-i}^{\infty} A_k, & j = L-K. \end{cases} \quad (3)$$

$$2. \quad K < i \leq L-K$$

$$p_{ij} = \begin{cases} 0, & 0 \leq j < i-K; \\ A_{K+j-i}, & i-K \leq j < L-K; \\ \sum_{k=L-i}^{\infty} A_k, & j = L-K. \end{cases} \quad (4)$$

Note that for the special case $L-K \leq K$ formula (3) for $0 \leq i \leq L-K$ is sufficient to determine p_{ij} and formula (4) is not needed.

Using the transition probabilities, we can numerically compute the stationary probabilities $\mathbf{\Pi} = (\pi_0, \pi_1, \dots, \pi_{L-K})$, where $\pi_i = \lim_{n \rightarrow \infty} P\{Q(n) = i\}, i \in \mathcal{Q}$.

4. Analytical evaluation of performance metrics

This section provides the analytical evaluation of performance indicators.

4.1. Message loss probability. For $Q(n) = k$ there are $L-k$ empty places in the buffer before the arrival of new messages. So all the extra incoming messages will be lost and the probability of any tagged message being lost is given by the following formula, assuming that $A(n) > 0$:

$$P_{Loss} = \frac{1}{1 - e^{-\lambda\Delta}} \sum_{k=0}^{L-K} \pi_k \sum_{i=1}^{\infty} \frac{i}{L-k+i} A_{L-k+i}. \quad (5)$$

There is no loss if the number of incoming messages is less or equal to $L-k$ and this metrics can be assessed as follows:

$$P_{NoLoss} = \frac{1}{1 - e^{-\lambda\Delta}} \sum_{k=0}^{L-K} \pi_k \sum_{j=1}^{L-k} A_j. \quad (6)$$

4.2. Waiting time distribution. The probability mass function of the random waiting time in slots before the transmission of a tagged message corresponds to a

message waiting time and is given by the formulas below:

$$\begin{aligned}
 f_W(i) = & \sum_{k=0}^{iK-1} \pi_k \sum_{m=1}^K \frac{m}{iK - k + m} A_{iK-k+m} + \\
 & + \sum_{k=0}^{iK-1} \pi_k \sum_{j=1}^{\infty} \frac{K}{(i+1)K - k + j} A_{(i+1)K-k+j} + \\
 & + \sum_{k=iK}^{\min((i+1)K, L-K)-1} \pi_k \sum_{m=1}^{(i+1)K-k} A_m + \\
 & + \sum_{k=iK}^{\min((i+1)K, L-K)-1} \pi_k \sum_{j=1}^{\infty} \frac{(i+1)K - k}{(i+1)K - k + j} A_{(i+1)K-k+j},
 \end{aligned} \tag{7}$$

where i represents the number of time slots the message waits in the queue. The above formula is valid for $i \in \{1, 2, \dots, \lceil \frac{L}{K} \rceil - 1\}$. If $i = 0$, the corresponding probability will be as shown below:

$$f_W(0) = \sum_{k=0}^{K-1} \pi_k \sum_{j=1}^{K-k} A_j + \sum_{k=0}^{K-1} \pi_k \sum_{j=1}^{\infty} \frac{K - k}{K - k + j} A_{K-k+j}. \tag{8}$$

4.3. Energy consumption per message transaction. For every transaction, the sequence of a sensor actions is as follows: a sensor searches the BS for constant time T_S with constant power P_{R_x} , randomly accesses the BS in constant time T_{Ac} with a random power P_{T_x} depending on the distance from the transmitter to the receiver, then with constant power P_{R_x} the sensor remains in the waiting mode for random time T_W and finally the transmission process lasts for constant time T_M with a random power level P_{T_x} . The needed energy during one data transaction is given by the following formula:

$$E_T = P_{R_x} \cdot T_{R_x} + P_{T_x} \cdot T_{T_x}, \quad \text{where } T_{R_x} = T_S + T_W \text{ and } T_{T_x} = T_{Ac} + T_M. \tag{9}$$

Here P_{R_x} and T_{R_x} are constant values, T_{R_x} and P_{T_x} are random variables with the probability density functions

$$f_{T_{R_x}}(t) = f_{T_W}(t - T_S), \quad f_{P_{T_x}}(y) = (2(y/A)^{2/\gamma})/\gamma R^2 y$$

correspondingly. The pdf of P_{T_x} corresponds to the radiochannel model $P_{T_x} = AD^{-\gamma}$, where A is the transmitted power and γ the path loss exponent. Making the convolution of $f_{T_{R_x}}(t)$ and $f_{P_{T_x}}(y)$ according to (9) we obtain the distribution $f_{E_T}(t, y)$ of the energy spent per transaction.

4.4. Distribution of sensor lifetime. During its lifetime a sensor switches between two modes: transaction mode and sleep mode (when there is no message to transmit). In the previous subsection, we explored the distribution of the energy used during a transaction period. Let $E_{Idle} = P_{Idle} \cdot T_{Idle}$ be the energy spent during a sleep period, where P_{Idle} is the constant power level used and T_{Idle} the exponentially distributed time with parameter λ_I , which determined by the sensor's data generation.

Now let us consider a cycle as consecutive transaction and sleep periods. Hence the energy for a cycle is $E_C = E_T + E_{Idle}$. Knowing $f_{E_T}(t, y)$ and $f_{E_{Idle}}(t)$, we get the pdf $f_{E_C}(t, y)$ of E_C as their convolution.

Let us denote $E(t)$ as a random variable of the energy used by the sensor at any given time t after the beginning of battery's lifetime. Let us also make a reasonable assumption that the sensor's batteries capacity U is much larger than the mathematical expectation of E_C ($U \gg E[E_C]$), and this means that after charging the batteries once, the sensor will perform a large number of cycles.

Considering the central limit theorem, $E(t)$ can be approximated by a normally distributed random variable with parameters that can be found through

$$\mu_{E(t)} = \frac{E[E_C]}{E[T_C]} t; \quad \sigma_{E(t)}^2 = \frac{\sigma^2[T_C]t}{(E[T_C])^3}. \quad (10)$$

Then we can get the distribution function of the sensor's lifetime T_L in the form of

$$\begin{aligned} F_{T_L}(t) &= P\{T_L < t\} = 1 - P\{T_L > t\} = \\ &= 1 - P\{U - E(t) > 0\} = 1 - P\{E(t) < U\} = \\ &= 1 - \Phi_{E(t)}(U), \text{ where } \Phi_{E(t)} \text{ is the cdf of } E(t). \end{aligned} \quad (11)$$

5. Conclusion

The work proposes an analytical model of interactions between IoT devices and a base station through narrowband-Internet of Things technology. The model is built as a queueing system and allows one to make an analytical estimation of performance metrics of the system. The next step in further research would be to add to this basic model some relay mobile vehicles, which could help to reduce the sensor's energy consumption and therefore increase its lifetime. Note that the proposed mathematical apparatus of embedded Markov chain for studying the above mentioned characteristics is applicable for Quality of Service indicators evaluation in tasks of analysis of efficiency of data exchange between drones in a swarm.

Acknowledgement

This paper has been supported by the RUDN University Strategic Academic Leader-

ship Program. The reported study was partially funded by RFBR, project number 20-07-01064.

REFERENCES

1. Varga P, Peto J, Franko A, Balla D, Haja D, Janky F, Soos G, Ficzere D, Maliosz M, Toka L. 5G support for Industrial IoT Applications— Challenges, Solutions, and Research gaps. Sensors. 2020; 20(3):828. <https://doi.org/10.3390/s20030828>.
2. Al-Absi, M.A., Al-Absi, A.A., Sain, M., Lee, H.J. (2020). A State of the Art: Future Possibility of 5G with IoT and Other Challenges. In: Pattnaik, P., Mohanty, S., Mohanty, S. (eds) Smart Healthcare Analytics in IoT Enabled Environment. Intelligent Systems Reference Library, vol 178. Springer, Cham.https://doi.org/10.1007/978-3-030-37551-5_3.
3. Muteba, K. F., Djouani, K., Olwal, T. (2022). 5G NB-IoT: Design, Considerations, Solutions and Challenges. Procedia Computer Science, 198, 86-93, <https://doi.org/10.1016/j.procs.2021.12.214>.
4. Iqbal, M., Abdullah, A. Y. M., Shabnam, F. (2020, June). An application based comparative study of LPWAN technologies for IoT environment. In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 1857-1860). IEEE, doi:10.1109/TENSYMP50017.2020.9230597.
5. Gbadamosi, S. A., Hancke, G. P., Abu-Mahfouz, A. M. (2020). Building upon NB-IoT networks: A roadmap towards 5G new radio networks. IEEE Access, 8, 188641-188672, doi:10.1109/ACCESS.2020.3030653.
6. Petrov V. et al. Vehicle-based relay assistance for opportunistic crowdsensing over narrowband IoT (NB-IoT) //IEEE Internet of Things journal. – 2017. – . 5. – №. 5. – . 3710-3723.
7. Moltchanov D., Koucheryavy Y. D-BMAP/D/1/K queuing system with priorities //Proceedings of the International Congress on Ultra Modern Telecommunications and Control Systems ICUMT'10, Moscow, October 18-20, 2010. – 2010. – . 1-5.
8. Grover R., Chaudhary H., Sharma G. Geo/G/1 System: Queues with Late and Early Arrivals //Intelligent Computing and Applications. – Springer, Singapore, 2021. – . 781-792.
9. Verdonck F., Bruneel H., Wittevrongel S. An All Geometric Discrete-Time Multiserver Queueing System // International Conference on Analytical and Stochastic Modeling Techniques and Applications. – Springer, Cham, 2019. – . 57-70.
10. Alfa A. S. Queueing theory for telecommunications: discrete time modelling of a single node system. – Springer Science Business Media, 2010.

УДК: 519.248

Расчет матрично-аналитической модели суперкомпьютера в переходном режиме

С.Н. Астафьев^{1,2}

¹ИПМИ КарНЦ РАН, ул. Пушкинская 11, Петрозаводск, Россия

²ИМИТ ФГБОУ ПетрГУ, пр. Ленина 33, Петрозаводск, Россия

seryumail@mail.ru

Аннотация

В работе рассмотрена задача расчета характеристик матрично-аналитической модели вычислительного кластера (суперкомпьютера) в переходном режиме. Для расчетов применен численный метод, основанный на методе Рунге-Кутты 4-го порядка решения обыкновенных дифференциальных уравнений и их систем. Найдена оценка для максимальной длины шага алгоритма.

Ключевые слова: Матрично-аналитический метод, переходный режим, метод Рунге-Кутты, обобщённый процесс рождения и гибели.

1. Введение

Стохастическое моделирование вычислительных систем позволяет установить их ключевые характеристики, выполнить анализ стационарности и оценить эффективность как при проектировании, так и в процессе их эксплуатации. Суперкомпьютеры являются одними из типов современных вычислительных систем, особенности которых делают моделирование особенно затруднительным [1].

Частным случаем стохастической модели суперкомпьютера является матрично-аналитическая модель суперкомпьютера с рандомизированной политикой переключения скоростей обслуживания. Анализ стационарного режима этой модели может быть найден в работах [2], [3], краткое описание этой модели приводится в разделе 2 настоящей статьи. Поведение такой системы может быть описано в виде обобщённого процесса рождения и гибели.

Обобщённый процесс рождения и гибели — это двухкомпонентный марковский процесс $X(t) = \{x(t), \phi(t)\}_{t \geq 0}$ с дискретным пространством состояний, где $x(t) \geq 0$ есть счётный уровень, который за один шаг может меняться не более чем на 1, а

Исследование выполнено за счет гранта Российского научного фонда № 21-71-10135,
<https://rscf.ru/project/21-71-10135/>

$\phi(t)$ — конечная *фаза** процесса в момент времени t . Такие процессы достаточно часто используются для исследования систем массового обслуживания M/M/-типа. При этом, под уровнем обычно понимается число заявок в системе, а под фазой — некоторое внутреннее состояние системы, влияющее на переходы цепи.

Для анализа такого процесса можно использовать метод, описанный в [4] и [5]. Этот метод подразумевает запись бесконечной генераторной матрицы обобщённого процесса рождения и гибели в блочном трёхдиагональном виде:

$$Q = \begin{pmatrix} A_0^{(0)} & A_0^{(+1)} & \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \dots \\ A_1^{(-1)} & A_1^{(0)} & A_1^{(+1)} & \ddots & \mathbb{O} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \dots \\ \mathbb{O} & A_2^{(-1)} & A_2^{(0)} & \ddots & \mathbb{O} & \mathbb{O} & \mathbb{O} & \mathbb{O} & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \dots & \dots & \dots \\ \mathbb{O} & \dots & \mathbb{O} & A_c^{(-1)} & A_c^{(0)} & A^{(+1)} & \mathbb{O} & \mathbb{O} & \dots \\ \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} & A^{(-1)} & A^{(0)} & A^{(+1)} & \mathbb{O} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \dots \end{pmatrix}, \quad (1)$$

где матрицы интенсивностей $A_*^{(+1)}$ описывают переходы на уровень выше, $A_*^{(-1)}$ — на уровень ниже, $A_*^{(0)}$ — описывают переходы внутри уровня, а матрицы \mathbb{O} являются нулевыми блоками. Все элементы вне главной диагонали являются неотрицательными, а суммы по строкам матрицы равны нулю.

Поведение системы в переходном режиме описывается системой уравнений Колмогорова–Чепмэна:

$$\frac{d\pi(t)}{dt} = \pi(t)Q, \quad (2)$$

$$\pi(0) = \pi_0,$$

где $\pi(t)$ — вектор, описывающий вероятностное распределение состояний системы в момент времени $t \geq 0$ при начальном состоянии π_0 .

2. Модель суперкомпьютера с рандомизированным переключением скорости обслуживания в переходном режиме

Рассмотрим систему массового обслуживания состоящую из c серверов. Система может работать в n разных режимах, со скоростями $f_1 < f_2 < \dots < f_n$ в режимах $1, 2, \dots, n$ соответственно.

Время между приходами заявок является экспоненциально распределённым с параметром λ . Приходящая в систему заявка может потребовать для своей

*Множества состояний фазы могут отличаться для малых уровней.

обработки i серверов[†] с вероятностью p_i , $i = 1, \dots, c$. Каждая заявка требует выполнения экспоненциально распределённого количества работы с параметром μ_i , где i - класс заявки.

При приходе заявки, система массового обслуживания может переключиться из j -того в k -тый скоростной режим с вероятностью $P_{j,k}^{(a)}$, где $j < k$. При уходе заявки, система может переключиться из j -того в k -тый режим с вероятностью $P_{j,k}^{(d)}$, где $j > k$. Матрицы $\mathbf{P}^{(a)}$ и $\mathbf{P}^{(d)}$ определяет структуру матриц $\mathbf{A}_*^{(+1)}$ и $\mathbf{A}_*^{(-1)}$ соответственно. Такую систему массового обслуживания назовём моделью суперкомпьютера с рандомизированным переключением скорости обслуживания.

При вычислении вероятностного распределения состояний модели суперкомпьютера, фаза рассматривалась в виде тройки $\langle m, s, q \rangle$, где m – номер скоростного режима, s – вектор длины c , который описывает число заявок каждого класса среди находящихся на обслуживании, q – класс заявки, находящейся в начале очереди. Блоки генераторной матрицы составлялись с помощью программы, рассматривающей все возможные переходы между уровнями и вычисляющей их интенсивности. Сравнение стационарных распределений, вычисленных для модели с модифицированным фазовым пространством, с результатами, полученными в [2] и [3], показало хорошее соответствие численных результатов.

3. Метод Рунге-Кутты 4-го порядка

Семейство методов Рунге-Кутты предназначено для численного решения задачи Коши для обыкновенных дифференциальных уравнений и их систем. Подробное описание метода может быть найдено в [6]. Для уравнения (2) одна итерация метода Рунге-Кутты 4-го порядка выглядит следующим образом:

$$\begin{aligned} \mathbf{k}_1 &= \boldsymbol{\pi}_{i-1} \mathbf{Q}, \\ \mathbf{k}_2 &= (\boldsymbol{\pi}_{i-1} + \mathbf{k}_1 \frac{h}{2}) \mathbf{Q}, \\ \mathbf{k}_3 &= (\boldsymbol{\pi}_{i-1} + \mathbf{k}_2 \frac{h}{2}) \mathbf{Q}, \\ \mathbf{k}_4 &= (\boldsymbol{\pi}_{i-1} + \mathbf{k}_3 h) \mathbf{Q}, \\ \boldsymbol{\pi}_i &= \boldsymbol{\pi}_{i-1} + \frac{h}{6} (\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4), \end{aligned} \tag{3}$$

где $\boldsymbol{\pi}_i \approx \boldsymbol{\pi}(h \cdot i)$, т.е. является приближённым решением уравнения (2) в точках $h \cdot i$. Пусть начальный вектор $\boldsymbol{\pi}_0$ при записи в блочном виде выглядит следующим образом:

$$\boldsymbol{\pi}_0 = (\boldsymbol{\pi}_0^{(0)}, \boldsymbol{\pi}_0^{(1)}, \boldsymbol{\pi}_0^{(2)}, \dots) = (\boldsymbol{\pi}_0^{(0)}, \boldsymbol{\pi}_0^{(1)}, \dots, \boldsymbol{\pi}_0^{(r)}, \mathbf{0}, \mathbf{0}, \dots), \tag{4}$$

[†]Число требуемых серверов будем называть *классом* заявки.

где векторы $\pi_0^{(i)}$ описывают начальное распределение уровней i , а $\mathbf{0}$ — нулевой вектор. При этом $\exists r : \forall i > r, \pi_0^{(i)} = \mathbf{0}$.

Рассмотрим операцию умножения вектора π вида (4) на матрицу \mathbf{Q} вида (1):

$$\begin{aligned} &(\pi^{(0)}, \pi^{(1)}, \dots, \pi^{(r)}, \mathbf{0}, \dots) \mathbf{Q} = \\ &(\pi^{(0)} \mathbf{A}_0^{(0)} + \pi^{(1)} \mathbf{A}_1^{(-1)}, \pi^{(0)} \mathbf{A}_0^{(+1)} + \pi^{(1)} \mathbf{A}_1^{(0)} + \pi^{(2)} \mathbf{A}_2^{(-1)}, \\ &\pi^{(1)} \mathbf{A}_1^{(+1)} + \pi^{(2)} \mathbf{A}_2^{(0)} + \pi^{(3)} \mathbf{A}_3^{(-1)}, \dots, \pi^{(c-1)} \mathbf{A}_{c-1}^{(+1)} + \pi^{(c)} \mathbf{A}_c^{(0)} + \pi^{(c+1)} \mathbf{A}^{(-1)}, \\ &\pi^{(c)} \mathbf{A}^{(+1)} + \pi^{(c+1)} \mathbf{A}^{(0)} + \pi^{(c+2)} \mathbf{A}^{(-1)}, \dots, \pi^{(r-2)} \mathbf{A}^{(+1)} + \pi^{(r-1)} \mathbf{A}^{(0)} + \pi^{(r)} \mathbf{A}^{(-1)}, \\ &\pi^{(r-1)} \mathbf{A}^{(+1)} + \pi^{(r)} \mathbf{A}^{(0)}, \pi^{(r)} \mathbf{A}^{(+1)}, \mathbf{0}, \mathbf{0}, \dots). \end{aligned}$$

При этой операции умножения все элементы результирующего вектора после уровня $r + 1$ равны нулю из-за структуры матрицы \mathbf{Q} и структуры вектора π . Таким образом один шаг метода Рунге-Кутты увеличивает число ненулевых уровней вектора π_i не более чем на 4. Используя это свойство, можно уменьшить хранимое число уровней вектора π_i до r на нулевом шаге алгоритма, а на каждом следующем увеличивать их число не более чем на 4.

4. Выбор длины шага алгоритма

Длина шага алгоритма $h > 0$ обычно выбирается исходя из необходимой погрешности численного метода. При решении уравнения (2) естественным образом возникают следующие два ограничения на решение:

- 1) $\pi_i \mathbf{1}^T = 1 \forall i$, где $\mathbf{1}$ – вектор из единиц. Это свойство выполняется при любой длине шага h , поскольку суммы по строкам матрицы \mathbf{Q} нулевые.
- 2) $\min(\pi_i) \geq 0 \forall i$. Это ограничение на решение явно запрещает вероятности нахождения в каком-либо состоянии становиться отрицательной.

Одна итерация метода Рунге-Кутты (3) может быть переписана (если подставить k_j и собрать коэффициенты по степеням \mathbf{Q}) следующим образом:

$$\pi_i = \pi_{i-1} (\mathbf{I} + h\mathbf{Q} + \frac{1}{2}(h\mathbf{Q})^2 + \frac{1}{6}(h\mathbf{Q})^3 + \frac{1}{24}(h\mathbf{Q})^4), \quad (5)$$

где \mathbf{I} – единичная матрица. Пусть $\alpha = \min(\mathbf{Q})$, из свойств матрицы \mathbf{Q} следует, что $\alpha < 0$, находится на главной диагонали и по модулю больше или равна любому другому элементу матрицы. Пусть $h \leq -\frac{1}{\alpha}$, тогда все элементы матрицы $\mathbf{C} = \mathbf{I} + h\mathbf{Q}$ неотрицательны, а (5) может быть переписана следующим образом:

$$\pi_i = \pi_{i-1} \left(\frac{3}{8}\mathbf{I} + \frac{1}{3}\mathbf{C} + \frac{1}{4}\mathbf{C}^2 + \frac{1}{24}\mathbf{C}^4 \right). \quad (6)$$

Следовательно, при выборе $h \leq -\frac{1}{\alpha}$ все элементы вектора π_i больше или равны нулю для любого i .

5. Численные результаты

В этом разделе в качестве примера, приводятся численные результаты для способа расчета, предложенного в 3-м разделе. В эксперименте использовались следующие параметры модели: $\lambda = 0.99$, $c = 2$, $f = (1, 2.2)$, $\mu = (1, 2)$, $p = (\frac{2}{3}, \frac{1}{3})$,

$$\mathbf{P}^{(a)} = \begin{pmatrix} 0.2 & 0.8 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{P}^{(d)} = \begin{pmatrix} 1 & 0 \\ 0.8 & 0.2 \end{pmatrix}. \quad (7)$$

Результаты для среднего числа клиентов представлены на Рисунке 1. Стационарное распределение было получено численно с помощью методов, описанных в [7].

Оценка локальной погрешности метода выполнялась по правилу Рунге:

$$\hat{\pi}(t_i + h) - \pi_{i+1,h} \approx \frac{1}{2^p - 1} (\pi_{i+1,h} - \pi_{i+1,2h}) = \mathbf{e}, \quad (8)$$

где $\hat{\pi}(t_i + h)$ точное решение в момент времени $t_i + h$, $\pi_{i+1,h}$ – численное решение в момент времени $t_i + h$, с шагом длины h ; $p = 4$ – порядок точности метода. Норма вектора оценки погрешностей $\|\mathbf{e}\|_1$ представлена на Рисунке 2.

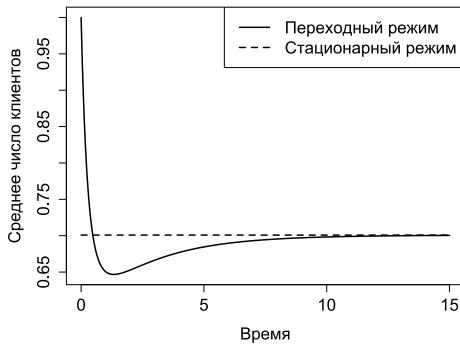


Рис. 1. Среднее число клиентов в переходном режиме.

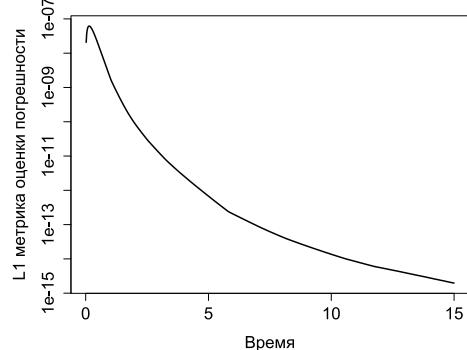


Рис. 2. ℓ_1 норма вектора оценки погрешности по правилу Рунге.

Длина шага h была выбрана равной 0.01, в начальный момент времени в системе 1 клиент: $\pi_0^{(1)} = (0.25, 0.25, 0.25, 0.25)$, $\pi_0^{(i)} = \mathbf{0} \forall i \neq 1$. Все вычисления производились в R версии 4.2.0.

По графикам можно сделать вывод о сходимости с течением времени метода к стационарному режиму и точности не менее 7 знаков при выбранных параметрах.

6. Заключение

В этой статье выполнена адаптация метода Рунге-Кутты к анализу переходного режима обобщённого процесса рождения и гибели на примере модели суперкомпьютера с рандомизированным переключением скоростей обслуживания. Приведённый метод позволяет численно рассчитать распределение переходного режима процесса, с линейным возрастанием памяти на каждом шаге алгоритма. Была найдена оценка для максимально возможной длины шага алгоритма, которую в будущем можно использовать в методе Рунге-Кутты с переменной длиной шага.

Литература

1. M. Harchol-Balter, Open problems in queueing theory inspired by datacenter computing, Queueing Systems 97 (1) (2021) 3–37. doi:10.1007/s11134-020-09684-6.
URL <https://doi.org/10.1007/s11134-020-09684-6>
2. R. M. Garimella, A. Rumyantsev, On an exact solution of the rate matrix of g/m/1-type markov process with small number of phases, Journal of Parallel and Distributed Computing 119 (2018) 172–178. doi:10.1016/j.jpdc.2018.04.013.
3. A. Rumyantsev, R. Basmadjian, A. Golovin, S. Astafiev, A three-level modelling approach for asynchronous speed scaling in high-performance data centres, in: Proceedings of the Twelfth ACM International Conference on Future Energy Systems, e-Energy '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 417–423. doi:10.1145/3447555.3466580.
URL <https://doi.org/10.1145/3447555.3466580>
4. Q.-M. He, Fundamentals of Matrix-Analytic Methods, Springer New York, New York, NY, 2014. doi:10.1007/978-1-4614-7330-5.
5. G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, Society for Industrial and Applied Mathematics, Philadelphia, 1999. doi:10.1137/1.9780898719734.
6. J. C. Butcher, Numerical Methods for Ordinary Differential Equations, 3rd Edition, Wiley, 2016.
7. D. Bini, G. Latouche, B. Meini, Numerical methods for structured Markov chains, Numerical mathematics and scientific computation, Oxford University Press, OCLC: ocm56807167.

UDC: 004.9:004.738.52

On Application of Source Code Analysis Techniques to HTML Pages Data Extraction

D.A. Orlov¹

¹National Research University "Moscow Power Engineering Institute", 14,
Krasnokazarmennaya str., Moscow, Russia

orlovdmal@mpei.ru

Abstract

Web scraping technique becomes more important as data grows in the Internet. There are lots of algorithms developed, most of them requires human assistance. The proposed approach using source-code analysis techniques for extracting data from HTML pages. The extracted information is divided into fields, which are parameters of extracted entities.

Keywords: Information retrieval, web scraping, data mining, source code analysis, pattern mining

1. Introduction

Currently, different commercial companies and public institutions are using web sites to provide different data. These data include technical documentation, information about consumer goods, address information and so-called open data. These data don't have to be structured and ready for automatic processing, and represented in different range of formats. But these data are often too large to be processed manually. This is why different approaches of automatic data extraction are developing rapidly [1],[2],[3]. In this paper we concentrate on data extraction from HTML web-sites, since it is most common way of data representation in the Internet. Most publications in this field are dedicated to data extraction by supervised machine learning methods application [4] or searching for data of known format (e.g., phone numbers and data) [5] or extracting data from pages with manually written templates [2],[3].

Nowadays, mostly used web scraping scenario is to use parsing utilities (such as ScraPy and BeautifulSoup) for HTML page parsing and heuristic algorithms or manually marked XPaths to locate specific parts of HTML page, containing required data.

To improve extraction quality web page rendering and image recognition are used. Machine learning in this case is used less often, since it requires manual web-page markup. Unsupervised learning is not achieved acceptable quality yet. Therefore, the task of unsupervised web scraping becomes more actual.

2. Data Extraction Task Formulation

Consider the following task. Let we have corpus of web-pages extracted from single web site. The pages have similar layout elements (header, menu, footer, etc.) and other parts (counters, captchas, scripts). Consider the corpus of web-pages contains large number of entities of interest which have (almost) the same format. The considered problem is to extract all relevant information from HTML code in form of entities and remove all unneeded HTML code.

In the simplest case consider each page contains an article which have text, header, date of publication and possibly some other attributes. In this case the article may be the entity of interest. In more complicated case the HTML page can contain table which contains characteristics of, e.g., mobile phones. In that case entity of interest is mobile phone, and the data of the entities are present in rows of the table.

In this research DOM tree analysis is used, since it is more accurate than regular expressions, and faster than rendering.

The idea of HTML extraction problem solving is to consider all information, repeated on several pages unimportant, since it likely belongs to elements of page design or counters. In that case each HTML page consists of common and unique parts (fig. 1).

Common part is the subtree of DOM tree, and unique part consists of several subtrees, each of them is child to common part subtree. In the most straightforward case common part contains the root of the DOM tree, but it is not necessary, since entities of interest can be contained in table rows, tables, div blocks etc.

Since HTML has formal grammar and can be processed by computer, it is possible to apply some programming language analysis techniques to it. Nowadays, program comprehension is the one of program analysis fields developing rapidly. These techniques make automatic refactoring possible. One of such techniques is programming idiom searching. Programming idioms are code fragments which occur in different software projects, and which solve one typical task [6], [7]. Entities of interest having common parts of their subtrees can be considered as programming idioms. Thus, applying programming idiom extraction techniques to HTML pages corpus will probably result in better HTML source segmentation.

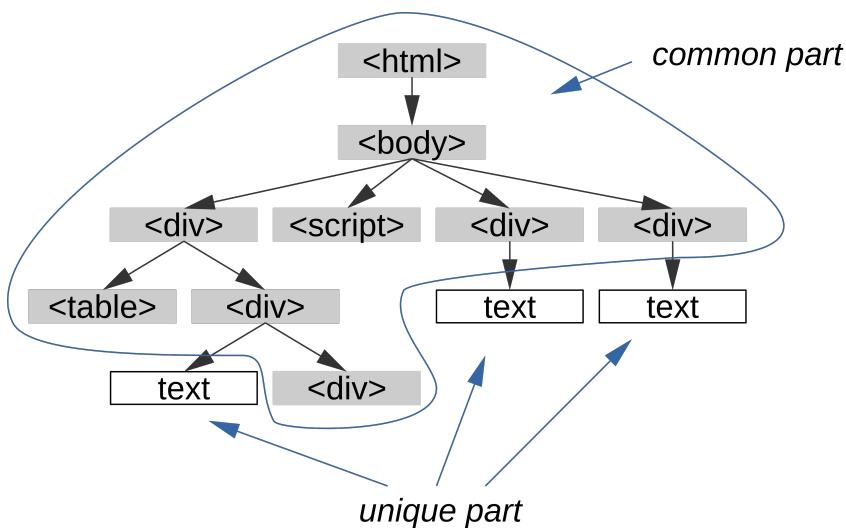


Fig. 1. Example of common and unique parts of HTML DOM tree

3. Suggested Approach

Consider programming idiom is the part of program abstract syntax tree (AST). Idiom arguments are subtrees, children of idiom subtree. There is difference between code clone and programming idiom. To become a clone, a snippet of code should appear in the program at least twice. The idiom, should appear more frequently and in different projects. On the other hand there is a difference between programming idiom finding task and frequent tree searching. The most frequent AST subtrees are usually short and uninformative. Thus, programming idiom is subtree of program AST which have meaning for programmer.

In [7] the programming idiom is defined as the subtree of program AST, which provides maximum compression of program AST in case the idiom is replaced by special node (e.g., frequently occurred program fragment is replaced by function). The metrics of AST compression used in the paper is based on Haelstad source code metric, which estimates information quantity in source code of the program.

Consider information quantity in original AST is:

$$H = N_1 \log_2(n_1) + N_2 \log_2(n_2) \quad (1)$$

where N_1 is number of operations in AST, n_1 is number of unique operations in AST, N_2 is number of operands in AST, n_2 is number of unique operands in AST. After program modification and replacing idiom with special node (e.g., replace repeated code fragments with function calls and adding function definition) number

operation and operands will change, so quantity if information will become (detailed explanation is given in [7]):

$$H' = (N_1 - CL_1 + C + L_1 + 1)\log_2(n_1) + (N_2 - CL_2 + C + L_2 + 1)\log_2(n_2) \quad (2)$$

where C is idiom occurrences count, L_1 - is number of operators in idiom subtree, L_2 - is number of operands in idiom subtree. Thus, the idiom efficiency metric is:

$$E = H - H' \quad (3)$$

The idiom extraction algorithm consists of two phases: tree database building and iterative idiom construction, at each step one node is added, thus finding local maximum of idiom efficiency.

To apply this formula to HTML DOM tree, consider all text nodes, tag parameter names and tag parameter values are operands, and all tag nodes are operations.

If DOM considered as AST, then programming idioms will become common parts of HTML pages, and idiom arguments can be considered as fields of extracted entities of interest. The results can be represented as JSON array.

Thus the data extraction algorithm is the following:

Step 1. For each HTML document in corpus.

Step 1.1. Transform HTML document into DOM tree.

Step 1.2. Add interested subtrees of DOM tree into tree database. List of interested subtrees depends on objects being searched (i.e., entities of interest).

During the experiment subtrees starting with tags "body", "tr", "table", "form" were added into database.

Step 2. Extract idioms from tree database.

Step 3. Select an idiom with maximum efficiency and print its arguments for all subtrees which contain the idiom.

4. Data Extraction Experiment

The proposed algorithm is implemented in Python3 language. For the experiment generated doxygen documentation for OpenCV programming library is used. The dataset contains 6893 HTML files, which has size 140 Mb.

The idiom with the highest efficiency metrics value occurs 4925 times, but there are idioms which have similar structure and more occurrences. Number of documents having similar structure, found via manually written regular expressions is 5229, thus recall of proposed algorithm is 0.941

The idiom extracted is presented on fig. 2. The idiom is automatically discovered template for processed HTML files. The idiom arguments - subtrees which contains variable parts are shown in bold.

```

<body $globalHandlers>
<div id="top">
  <div id="titlearea"><table><tbody><tr style="height: 56px;">
    <td id="projectlogo"></td>
    <td style="padding-left: 0.5em;">
      <div id="projectname">OpenCV</div>
      <div id="projectbrief">Open Source Computer Vision</div>
    </td>
  </tr></tbody></table></div>
  <script type="text/javascript">var searchBox = new SearchBox("searchBox",
"../../search",false,'Search');</script>
<script src="../../menudata.js" type="text/javascript" />
<script src="../../menu.js" type="text/javascript" />
<script type="text/javascript">$(function() {
initMenu('../../',true,false,'search.php','Search'); $(document).ready(function()
{ initSearch(); });});</script>
<div id="main-nav" />
<div id="MSearchSelectWindow" onkeydown="return searchBox.OnSearchSelectKey(event)"
onmouseout="return searchBox.OnSearchSelectHide()" onmouseover="return
searchBox.OnSearchSelectShow()" />
<div id="MSearchResultsWindow"><iframe src="javascript:void(0)" name=MSearchResults
id=MSearchResults" /></div>
<div class="navpath" id="nav-path"><ul>
  <li class="navelem"><a href=$str class=el>$str</a></li>
..</ul></div>
</div>
<div class="header">$body</div>
<div class="contents">$body</div>
<hr class="footer" />
<address class="footer"><small>$str
  <a href="http://www.doxxygen.org/index.html"><img alt="doxygen src=../../doxygen.png"
class="footer" /></a> 1.8.13
</small></address>
<script type="text/javascript">//![CDATA[ addTutorialsButtons(); //]]></script>
</body>

```

Fig. 2. Example of extracted idiom

As we can see, the algorithm extracted article content and header as well as other distinct parts (e.g. document generation time contains in footer). Thus the possible directions of algorithm improvement are recall improvement, and idiom argument filtration.

5. Conclusion

This paper is dedicated to application program analysis methods to data extraction from HTML pages. The program comprehension algorithms now make possible more accurate analysis of HTML pages corpora. Proposed and implemented algorithm of unsupervised data extraction from HTML pages. The algorithm is tested on extracting articles from automatically generated documentation.

REFERENCES

1. Haddaway, N. R. (2015). The use of web-scraping software in searching for grey literature. Grey J, 11(3), 186-90.
2. Diouf, R., Sarr, E. N., Sall, O., Birregah, B., Bousso, M., & Mbaye, S. N. (2019, December). Web scraping: state-of-the-art and areas of application. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 6040-6042). IEEE.

3. Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-based systems*, 70, 301-323.
4. Karthikeyan, T., Sekaran, K., Ranjith, D., & Balajee, J. M. (2019). Personalized content extraction and text classification using effective web scraping techniques. *International Journal of Web Portals (IJWP)*, 11(2), 41-52.
5. Gupta, S., Kaiser, G., Neistadt, D., & Grimm, P. (2003, May). DOM-based content extraction of HTML documents. In *Proceedings of the 12th international conference on World Wide Web* (pp. 207-214).
6. M. Allamanis, C. Sutton, Mining Idioms from Source Code // FSE 2014 Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering, pp. 472-483, Hong Kong, China — November 16 - 21, 2014
7. Orlov D.A. An algorithm of idiom search in program source codes using subtree counting // Software & Systems, 2022, vol. 35, no. 1, pp. 065–074 https://doi.org/10.1007/978-3-030-61470-6_4

UDC: 519.872

Queueing system for analyzing the operation of 5G network with NS under preemption-based scheduler

K.Y.B. Adou¹, E.V. Markova¹, Yu.V. Gaidamaka^{1,2}

¹Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

²Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov St, Moscow, 119333, Russian Federation

{1042205051, markova-ev, Gaidamaka-yuv}@rudn.ru

Abstract

The rapid deployment of the fifth generation (5G) of wireless systems since 2020 is profoundly transforming the telecommunication area, paving the way for new-generation networks that include more devices, are more reliable and enable faster communications. The network slicing (NS) technology, considered as one of the 5G systems' key features, allows the operation of several virtual networks called network slice instances (NSIs) on top of the same physical infrastructure, e.g. a base station. Ensuring a satisfactory quality of service (QoS) level under NS technology is an important task, probably requiring to use the guaranteed's bit rate (GBR) QoS concept. In this paper, a queueing system model for analyzing the operation of three 5G NSIs under a preemption-based scheduler maintaining the required QoS levels is proposed. The system's key performance indicators (KPIs) are defined and computed.

Keywords: 5G network, NS, preemption, scheduling, NSI, iterative method, KPI, queueing theory, network capacity, GBR, wireless system, service requirement.

1. Introduction

Recently, the network slicing (NS) entered the phase of commercialization around the globe and, thus, is drawing the attention of researchers and scientists [1]. In all the 5G network's systems components, the NS technology is considered as one

This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipient Adou K.Y.B., Methodology, Formal analysis, Software, Validation, Writing). The reported study was funded by RFBR, project number 20-07-01052 (recipient Markova E.V., Validation, Visualization, Supervision, Writing). The reported study was funded by RFBR, project number 20-07-01064 (recipient Gaidamaka Yu.V., Conceptualization, Validation, Visualization, Supervision).

of the most promising [2]. In practice, NS offers flexible solutions to manage/share efficiently the physical network resources through logical platforms known as network slice instances (NSIs), customizable and adaptable to the very specific needs of all sorts of tenants [1, 2, 3].

Since the NSIs' tenants need to provide a certain quality of service (QoS) level to their users, NS sustaining the various user's types with the required QoS levels may raise a problem whose solution lies in the guaranteed's bit rate (GBR) concept within the service's level agreement (SLA) framework [2]. Moreover, adding to NS systems preemption schedulers based on the 3GPP priority's levels range may be of significance for meeting the QoS requirements [1, 4].

In this paper, one constructs and analyzes a queueing system illustrating the functioning of three custom NSIs on top of one 5G base station (BS) under a preemption-based scheduler with consideration of the GBR SLA concept.

2. System model

Let us consider an internet provider renting out its fifth generation (5G) base station (BS) to three tenants based on the network slicing (NS) technology, i.e. each tenant is assigned a unique fully customizable and logical wireless network — a network slice instance (NSI). Let C be the 5G BS's total network capacity and C_s — the s -th NSI's overall network capacity, $s \in \{1, 2, 3\}$, under the condition $C_1 + C_2 + C_3 \geq C$. We consider that the internet provider guarantees a minimum network capacity to each tenant [5, 6], i.e. each tenant is assigned a guaranteed network capacity included in its overall network capacity. Let Q_s be the s -th NSI's guaranteed network capacity under the condition $Q_1 + Q_2 + Q_3 \leq C$.

Let us consider the Poisson arrival process of a unique call type at the s -th NSI with rate λ_s , $s \in \{1, 2, 3\}$, i.e. $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$. The arriving call requires b_s amount of resources for starting service under the condition $b_s \leq Q_s$. Let us consider the case when the service requirements b_s are uniform, i.e. $b_1 = b_2 = b_3 := b$, $b \in \mathbb{R}^+$. The average service time is exponentially distributed with the mean μ_s^{-1} , i.e. $\boldsymbol{\mu} = (\mu_1^{-1}, \mu_2^{-1}, \mu_3^{-1})$.

A summary of the system's additional notations is provided in the table 1.

The radio admission control (RAC) scheme for accessing the s -th NSI is organized so that when the number n_s of servicing calls is less than $\lfloor Q_s/b \rfloor$, $s \in \{1, 2, 3\}$, and the amount of available resources at the BS is less than b , a preemption of one servicing call occurs at the \hat{s} -th NSI, $\hat{s} \in \{1, 2, 3\} \setminus \{s\}$, once a new call arrives at that NSI. Let us assign priority levels to the custom NSIs for clarifying the preemption method. Let the highest priority level "1" be assigned to the 1st NSI, the medium "2" — to the 2nd NSI and the lowest "3" — to the 3rd NSI.

Notation	Description
$\lfloor C_s/b_s \rfloor$	The maximum number of calls that may be admitted for service with the s -th NSI's overall network capacity, $\mathbf{N}^{\max} = (\lfloor C_1/b_1 \rfloor, \lfloor C_2/b_2 \rfloor, \lfloor C_3/b_3 \rfloor)$
$\lfloor Q_s/b_s \rfloor$	The maximum number of calls that may be admitted for service with the s -th NSI's guaranteed network capacity, $\mathbf{N}^g = (\lfloor Q_1/b_1 \rfloor, \lfloor Q_2/b_2 \rfloor, \lfloor Q_3/b_3 \rfloor)$
n_s	The current number of calls at the s -th NSI, $\mathbf{n} = (n_1, n_2, n_3)$
\mathbf{e}_s	The s -th row of the identity matrix of size 3×3
\mathbf{J}	The three-dimensional all-ones vector

Table 1. The system's additional notations

One considers that upon arrival at its custom NSI a new call is bound to one path — immediate admission, delayed admission or blocking. Let us say the new call's admission is immediate when the number n_s of servicing calls at its custom NSI is less than $\lfloor C_s/b \rfloor$ and the amount of available resources at the BS is greater than or equal to b , i.e. $(\mathbf{n} + \mathbf{e}_s - \mathbf{N}^{\max}) \cdot \mathbf{e}_s \leq 0 \wedge (\mathbf{n} + \mathbf{e}_s) \cdot b\mathbf{J} \leq C, s \in \{1, 2, 3\}$. Let us say the call's admission is delayed due to the preemption's method application when the number n_s of servicing calls at its NSI is less than $\lfloor Q_s/b \rfloor$ and the amount of available resources at the BS is less than b , i.e. $(\mathbf{n} + \mathbf{e}_s - \mathbf{N}^g) \cdot \mathbf{e}_s \leq 0 \wedge (\mathbf{n} + \mathbf{e}_s - \mathbf{e}_{\hat{s}}) \cdot b\mathbf{J} \leq C, \hat{s} = \max \{\check{s} \in \{1, 2, 3\} : (\mathbf{n} - \mathbf{N}^g) \cdot \mathbf{e}_{\check{s}} > 0\}$. Lastly, let us say the new call is blocked when the number n_s of servicing calls at its NSI is either equal to $\lfloor Q_s/b \rfloor$ and the amount of available resources at the BS is less than b , or simply equal to $\lfloor C_s/b \rfloor$.

3. Mathematical model

According to the defined RAC scheme and considering the Poisson distributed arrival processes, plus the exponentially distributed service times, one may describe the system's behavior using a three-dimensional Markov process $\mathbf{X}(t) = \{X_1(t), X_2(t), X_3(t), t > 0\}$, where $X_s(t)$ — the number of servicing calls at the s -th NSI at time t , $s \in \{1, 2, 3\}$, over the system's state space:

$$\Omega = \{\mathbf{n} \in \mathbb{N}^3 : (\mathbf{n} - \mathbf{N}^{\max}) \cdot \mathbf{J} \leq 0 \wedge \mathbf{n} \cdot b\mathbf{J} \leq C\}, \quad (1)$$

where \mathbb{N}^3 — the set of all three-dimensional vectors with natural elements.

Let us introduce the main subsets [7] of the system's state space Ω for the model's further investigation:

- $\Omega_s^{\text{cap}}, s \in \{1, 2, 3\}$ — the subset with the system's states, where the admission of a new arriving call at the s -th NSI is delayed due to the preemption of one servicing call at the \hat{s} -th NSI, $\hat{s} \in \{1, 2, 3\} \setminus \{s\}$:

$$\Omega_s^{\text{cap}} = \{\mathbf{n} \in \Omega : (\mathbf{n} - \mathbf{N}^g) \cdot \mathbf{e}_s < 0 \wedge (\mathbf{n} + \mathbf{e}_s) \cdot b\mathbf{J} > C\}, \quad (2)$$

- $\Omega_s^{\text{iad}}, s \in \{1, 2, 3\}$ — the subset with the system's states, where the admission of a new arriving call at the s -th NSI is immediate:

$$\Omega_s^{\text{iad}} = \{\mathbf{n} \in \Omega : (\mathbf{n} - \mathbf{N}^{\max}) \cdot \mathbf{e}_s < 0 \wedge (\mathbf{n} + \mathbf{e}_s) \cdot b\mathbf{J} \leq C\}. \quad (3)$$

Let us illustrate the logical relations between the system's main subsets (2), (3) with the Venn's diagram style. Fig. 1 represents the Venn diagram of the system's main subsets, e.g. the subsets' intersection $\Omega_1^{\text{iad}} \cap \Omega_2^{\text{iad}} \cap \Omega_3^{\text{iad}}$ regroups the system's states, where the admission of any new arriving call at the system is immediate.

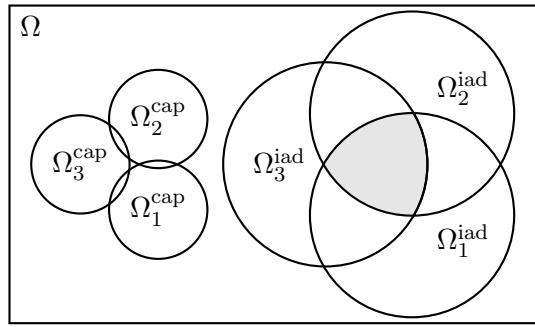


Fig. 1. The Venn diagram of the system's main subsets.

One describes the discussed Markov process of the system's state space Ω with the system of equilibrium equations:

$$\begin{aligned} P(\mathbf{n}) \left(\lambda \cdot \sum_{s=1,2,3} I_{\Omega_s^{\text{iad}} \cup \Omega_s^{\text{cap}}}(\mathbf{n}) \mathbf{e}_s + \mathbf{n} \cdot \boldsymbol{\mu} \right) = \\ \lambda \cdot \sum_{s=1,2,3} (P(\mathbf{n} - \mathbf{e}_s) H(\mathbf{n} \cdot \mathbf{e}_s) + P(\mathbf{n} - \mathbf{e}_s + \mathbf{e}_{\hat{s}}) I_{\Omega_s^{\text{cap}}}(\mathbf{n} - \mathbf{e}_s + \mathbf{e}_{\hat{s}})) \mathbf{e}_s + \\ \boldsymbol{\mu} \cdot \sum_{s=1,2,3} (P(\mathbf{n} + \mathbf{e}_s) I_{\Omega_s^{\text{iad}}}(\mathbf{n}) (\mathbf{n} + \mathbf{e}_s) \cdot \mathbf{e}_s) \mathbf{e}_s, \quad (4) \end{aligned}$$

where $\hat{s} = \max \{\check{s} \in \{1, 2, 3\} : (\mathbf{n} - \mathbf{N}^g) \cdot \mathbf{e}_{\check{s}} > 0\}$, $P(\mathbf{n})$, $\mathbf{n} \in \Omega$ — the stationary probability that the system is in the state \mathbf{n} , $H(*)$ — the Heaviside function, and $I_o(*)$ — the Indicator function.

One notes that the Markov process describing the system's behavior is not reversible. Therefore, one may compute the system's stationary probability distribution

$\mathbf{P} = (P(\mathbf{n}))$, $\mathbf{n} \in \Omega$, i.e. a $|\Omega| \times 1$ matrix, using an iterative method [8] for solving the system of equilibrium equations:

$$\mathbf{A}^\top \mathbf{P} = \mathbf{0}, \quad (5)$$

where \mathbf{A} — the infinitesimal generator of Markov process, i.e. a $|\Omega| \times |\Omega|$ matrix, whose entries $A(\mathbf{n} \in \Omega, \hat{\mathbf{n}} \in \Omega)$ are computed as follows: when $\mathbf{n} \neq \hat{\mathbf{n}}$

$$A(\mathbf{n}, \hat{\mathbf{n}}) = \begin{cases} \boldsymbol{\lambda} \cdot \mathbf{e}_s, & \text{if } \hat{\mathbf{n}} = \mathbf{n} + \mathbf{e}_s, \quad \text{s.t. } \mathbf{n} \in \Omega_s^{\text{iad}}, \\ & \text{else if } \hat{\mathbf{n}} = \mathbf{n} + \mathbf{e}_s - \mathbf{e}_{\hat{s}}, \quad \text{s.t. } \mathbf{n} \in \Omega_s^{\text{cap}}, \\ (\mathbf{n} \odot \boldsymbol{\mu}) \cdot \mathbf{e}_s, & \text{if } \hat{\mathbf{n}} = \mathbf{n} - \mathbf{e}_s, \quad \text{s.t. } \hat{\mathbf{n}} \in \Omega, \\ 0, & \text{otherwise,} \quad \text{i.e. } \hat{\mathbf{n}} \in \Omega \setminus \{\mathbf{n}\}, \end{cases} \quad s \in \{1, 2, 3\}, \quad \hat{s} = \max \{\check{s} \in \{1, 2, 3\} : (\mathbf{n} - \mathbf{N}^g) \cdot \mathbf{e}_{\check{s}} > 0\}, \quad (6a)$$

when $\mathbf{n} = \hat{\mathbf{n}}$

$$A(\mathbf{n}, \mathbf{n}) = - \sum_{\hat{\mathbf{n}} \in \Omega \setminus \{\mathbf{n}\}} A(\mathbf{n}, \hat{\mathbf{n}}). \quad (6b)$$

After obtaining the system's stationary probability distribution \mathbf{P} , one may compute its key performance indicators (KPIs). Let us propose the following KPIs for the model's further investigation:

- The mean number of servicing calls at the s -th custom NSI, $s \in \{1, 2, 3\}$, and at the system

$$\sum_{\mathbf{n} \in \Omega} P(\mathbf{n}) \mathbf{n} \cdot \mathbf{e}_s; \quad \sum_{\mathbf{n} \in \Omega} P(\mathbf{n}) \mathbf{n} \cdot \mathbf{J}; \quad (7)$$

- The immediate admission probability of a new arriving call at the s -th NSI, $s \in \{1, 2, 3\}$, and at the system

$$\sum_{\mathbf{n} \in \Omega_s^{\text{iad}}} P(\mathbf{n}); \quad \sum_{\substack{\mathbf{n} \in \bigcap_{s=1,2,3} \Omega_s^{\text{iad}}}} P(\mathbf{n}); \quad (8)$$

- The delayed admission probability of a new arriving call at the s -th NSI, $s \in \{1, 2, 3\}$, due to the preemption's method application

$$\sum_{\mathbf{n} \in \Omega_s^{\text{cap}}} P(\mathbf{n}); \quad (9)$$

- The blocking probability of a new arriving call at the s -th NSI, $s \in \{1, 2, 3\}$, and at the system

$$\sum_{\mathbf{n} \in \Omega \setminus (\Omega_s^{\text{iad}} \cup \Omega_s^{\text{cap}})} P(\mathbf{n}); \quad \sum_{\mathbf{n} \in \bigcap_{s=1,2,3} \Omega \setminus (\Omega_s^{\text{iad}} \cup \Omega_s^{\text{cap}})} P(\mathbf{n}). \quad (10)$$

4. Conclusion

One proposed a queueing system model for analyzing the operation of three 5G network slice instances (NSIs) under a preemption-based scheduler, maintaining the required QoS levels. The model's general specifics were presented. The system's key performance indicators (KPIs) were defined and computed.

REFERENCES

1. J. M. Meredith, F. Firmin, M. Pope, Release 16 Description; Summary of Rel-16 Work Items, Technical report (TR) 21.916, 3rd Generation Partnership Project (3GPP), version 16.1.0 (01 2022).
2. A. Sultan, M. Pope, Feasibility study on new services and markets technology enablers for network operation; Stage 1, Technical report (TR) 22.864, 3rd Generation Partnership Project (3GPP), version 15.0.0 (09 2016).
3. 5G industry campus network deployment guideline, Official Document NG.123, GSM Association (GSMA), version 2.0 (10 2021).
4. J. M. Meredith, M. C. Soveri, M. Pope, Management and orchestration; 5G end to end Key Performance Indicators (KPI), Technical specification (TS) 28.554, 3rd Generation Partnership Project (3GPP), version 17.6.0 (03 2022).
5. G. P. Basharin, Yu. V. Gaidamaka, K. E. Samouylov, Mathematical Theory of Teletraffic and its Application to the Analysis of Multiservice Communication of Next Generation Networks, Automatic Control and Computer Sciences 47 (2) (2013) 62–69. doi:10.3103/S0146411613020028.
6. N. Yarkina, L. M. Correia, D. Moltchanov, Y. Gaidamaka, K. Samouylov, Multi-tenant resource sharing with equitable-priority-based performance isolation of slices for 5g cellular systems, Computer Communications 188 (2022) 39–51. doi:10.1016/j.comcom.2022.02.019.
7. K. Devlin, The Joy of Sets, Undergraduate texts in mathematics, Springer New York, New York, NY, 1993. doi:10.1007/978-1-4612-0903-4.
8. S. N. Stepanov, Theory of Teletraffic: Concepts, Models, Applications [Teoriya teletraffika: kontseptsii, modeli, prilozheniya], Goryachaya Liniya-Telekom, Moscow, 2015, in Russian.

UDC: 519.872

Existence of stationary queue-size distributions in the systems that work only on the biggest batches of customers

R.V. Razumchik¹, L.A. Meykhanadzhyan², D.A. Pyatkina²

¹Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, Moscow, Russian Federation

²Financial University under the Government of the Russian Federation, Moscow,
Russian Federation

rrazumchik@ipiran.ru, lamejkhanadzhyan@fa.ru, dapiatkina@fa.ru

Abstract

Consideration is given to the infinite capacity FIFO queueing systems with a single server, that work only on the biggest batches of customers. New customers arrive only in batches. Upon arrival of a batch its size is compared with the current total number of customers the system. If the size of the batch is larger than that number, all customers residing in the system (including the one in server) are pushed-out and the arrived batch enters the system; otherwise the new batch is considered as lost. Conditions of the existence of the stationary queue-size distributions of such systems are of interest. It is shown that in the classical cases they follow the intuition. Yet the preliminary analysis shows, that finiteness of the mean batch size may not be necessary for the stability of such queues.

Keywords: batch arrivals, queue skipping policy, stability

1. Introduction

The purpose of this short note is to state some new problems (and give only a few answers) concerning the stability of one specific type of queueing systems — queueing systems, which work only on the biggest batches of customers. Apparently they were firstly introduced in [1] and since then some efforts have been made (see, for example, [2, 3]) to obtain deeper insights into their properties. In order to work only on the biggest batch of customers a queueing system adopts the so-called queue skipping policy. Roughly speaking (see the detailed description in the next section), if the size of the arriving batch of customers is smaller than the total number of customers currently present in the system, then the arriving batch skips the queue

The reported study was funded by RFBR, project number 20-07-00804.

and leaves the system without having further any effect on it; otherwise, the arriving batch empties out the system and occupies the system itself.

The analytic analysis of queues with such a queue skipping policy under more or less general assumptions about the arrival/service times and the batch-size distribution is a challenge. Even in the markovian case (and i.i.d. batch-sizes) the stationary distribution of the queue-size is in a certain sense intractable (see [1, Theorem 1]): computation of the system's idle probability involves evaluation of the infinite product and it is not clear when to stop computations especially if the batch-size distribution is heavy-tailed. For two and more (infinite-capacity) queues with one dedicated flow, which are interconnected with the considered queue skipping policy, the joint stationary distributions of the queues' content are not available in the literature. Moreover, even the stability conditions (i.e. the conditions of the existence of the stationary queue-size distributions) remain an open issue. This latter issue is addressed to some extent further in this short note.

2. System's description and the problem statement

Consider an infinite capacity FIFO queue with a single server. Customers arrive to the system in batches and the inter-arrival times A_1, A_2, \dots are i.i.d. with the common distribution function $F(x) = P(A < x)$, having mean EA . The size of an arriving batch becomes known upon its arrival and is the random variable with the given probability distribution $\{g_n, n \geq 1\}$, having mean $EG = \sum_{n=1}^{\infty} n g_n$. The following queue skipping policy is adopted in the system. Whenever a batch arrives to the system its size, say \hat{G} , is compared with the remaining total number of customers in the system, say \tilde{G} . If $\hat{G} > \tilde{G}$, then all customers, which are currently available in the system, are instantly removed from it, and the whole batch \hat{G} is placed in the queue and the first customer in the batch enters server. If $\hat{G} \leq \tilde{G}$ the new batch leaves the system without having any effect on it. Whenever the server becomes free one customer from the queue (if there is any) enters server. Service times S_1, S_2, \dots are i.i.d. with the common distribution function $B(x) = P(S < x)$, having mean ES .

Denote by $\xi(t)$ the total number of customers in the system at time t . Let $p_n = \lim_{t \rightarrow \infty} P(\xi(t) = n)$, $n \geq 0$. The problem is to find the conditions under which the system is stable i.e. the conditions for the existence of the stationary distribution $\{p_n, n \geq 0\}$.

Note that under the considered queue skipping policy new arrivals may push out customers in the queue. Yet such systems cannot be reduced to queues with negative arrivals/signals [4, 5] and ordered entry queues (see, for example, [6, 7, 8]).

3. Stability criterions

Under natural conditions the considered type of (infinite capacity) queueing systems posses the classical regeneration property. This argument holds also for interconnected (infinite capacity) queues, when each queue is equipped with the considered skipping policy. Already in [1] it was shown that if A and S are exponentially distributed and the batch-size distribution $\{g_n, n \geq 1\}$ is geometric, then the queue, working in isolation, is unconditionally stable. But already for the general batch-size distribution the stability condition is not provided in [1]. In order to apply the Forster's theorem, the exact form of $\{p_n, n \geq 0\}$ is required; yet in the most interesting cases it will not be available.

The difficulty of stability analysis of the considered type of queues stems from the observation that a direct proof of the finiteness of the mean regeneration period is prohibitively complicated. But it turns out that this difficulty can be circumvented (to some extent) by changing the point of view and searching for stability conditions in terms of the drift. This approach for stability analysis of regenerative queueing systems is summarized in the recent monograph [9]. Its key ingredient is the drift δ_n , defined as $\delta_n = \mathbb{E}(\xi_{i+1} - \xi_i | \xi_i = n)$, $n \geq 0$, where ξ_i is the value (on the i th transition) of an irreducible aperiodic Markov chain embedded into $\{\xi(t), t \geq 0\}$. If $|\delta_n| < \infty$ for all n and $\lim_{n \rightarrow \infty} \sup \delta_n < 0$, then $\{\xi(t), t \geq 0\}$ is ergodic (see [10]).

Consider an exponential queueing system with the arrival rate $(\mathbb{E}A)^{-1}$ and the service rate $(\mathbb{E}S)^{-1}$, and the general batch-size distribution $\{g_n, n \geq 1\}$. If ξ_i is the number of customers in the system upon the i th arrival, then it can be shown that

$$|\delta_n| \leq n \left(\left(2 \frac{(\mathbb{E}S)^{-1}}{(\mathbb{E}A)^{-1} + (\mathbb{E}S)^{-1}} \right)^n - 1 \right) \sum_{j=n+1}^{\infty} g_j + n \sum_{j=0}^n g_j + 2 \sum_{j=n+1}^{\infty} j g_j.$$

Thus for the system's stability it is necessary that the mean batch size $\mathbb{E}G$ is finite. Further analysis shows that this condition is also sufficient, which is in perfect alignment with the intuition. This result also holds for i.i.d. inter-arrival times under the additional condition that their mean $\mathbb{E}A$ is finite. Specifically, it can be shown that

$$\delta_n \leq -\text{"mean number of customers served during one transition from } n\text{"} + \sum_{j=n+1}^{\infty} j g_j,$$

and, eventually, that $\lim_{n \rightarrow \infty} \sup \delta_n < 0$.

Even though for Poisson arrivals and i.i.d. service times ξ_i is simply the number of customers in the system upon the i th service completion (but not departure), the analysis becomes more involved (and follows the different lines). The readily

available sufficient stability condition is $ESEG < \infty$, which again is being supported by the intuition. Yet the relation for δ_n shows that finiteness of the mean batch size (and the mean service time) may not be necessary for the system's stability. Preliminary analysis shows that the condition $ESEG < \infty$ can be replaced by the weaker one:

$$\sum_{l=1}^{\infty} \frac{\sqrt{\sigma_1 \times \cdots \times \sigma_l}}{\sum_{j=l+1}^{\infty} g_j} < \infty, \text{ where } \sigma_i = \sum_{j=1}^i \left(\frac{g_i}{\sum_{k=j+1}^{\infty} g_j} \right)^2.$$

Although it is not clear how the latter inequality can be checked in practice, the distribution $g_n = \frac{1}{n(n+1)}$, $n \geq 1$, shows that at least one case, when the inequality holds, does exist.

Generalization to the i.i.d. inter-arrival and service times (and the general batch size distribution) remains an open question. The main conjecture here is that the assumptions of the queue skipping policy are so strong, that a queue must be stable with the minimum requirements. The adopted framework seems to be suitable for the stability analysis of interconnected queues with the queue skipping policy. Here the argumentation suggested in [11] for the analysis of multi-class retrial queues seems to be the proper angle of attack.

REFERENCES

1. Marin A., Rossi S. A queueing model that works only on biggest jobs // European Workshop on Performance Engineering. Ed. M. Gribaudo, M. Iacono, T. Phung-Duc, R. Razumchik. Lecture Notes in Computer Science ser. Springer. 2020. V. 12039. P. 118–132.
2. Matyushenko S. I., Razumchik R. V. Stationary characteristics of discrete-time $Geo/G/1$ queue with batch arrivals and one queue skipping policy // Inform. Primen., 2020. V. 14. No. 4. P. 25–32.
3. Zeifman A.I., Razumchik R.V., Satin Y.A., Kovalev I.A. Ergodicity bounds for the Markovian queue with time-varying transition intensities, batch arrivals and one queue skipping policy // Applied Mathematics and Computation. 2021. V.395. Art. 125846.
4. Dudin A.N., Klimenok V.I., Vishnevsky V.M. The theory of queuing systems with correlated flows. Heidelberg, Germany: Springer, 2019. 447 p.
5. Gelenbe E. R'eseaux stochastiques ouverts avec clients negatifs and positifs, et reseaux neuronaux // Comptes-Rendus de l'Academie des Sciences. 1989. V. 309. Serie II. P. 972–982.

6. Disney R. L. Some multichannel queueing problems and ordered entry // J. Ind. Eng. 1962. V. 13. P. 46–48.
7. Nanwijn W. M. A note on many-server queueing system with ordered entry, with an application to conveyor theory // J. Appl. Prob. 1983. V. 20. No. 1. P. 144–152.
8. Zaryadov I. S., Meykhanadzhyan L. A., Milovanova T. A., Razumchik R. V. On the method of calculating the stationary distribution in the finite two-channel system with ordered input // Systems and Means of informatics. 2015. V. 25. No. 3. P. 44–59.
9. Morozov E., Steyaert B. Stability analysis of regenerative queueing models. Cham, Switzerland: Springer, 2021. 195 p.
10. Pakes A. G. Some Conditions for ergodicity and recurrence of Markov chains // Operations Research. 1969. V. 17. P. 1058–1061.
11. Avrachenkov K., Morozov E., Nekrasova R. Stability analysis of two-class retrial systems with constant retrial rates and general service times // arXiv. 2021. Available at: <https://arxiv.org/abs/2110.09840> (accessed May 26, 2022).

UDC: 004.852

A machine learning approach for predicting SINR

E. V. Bobrikova¹, A. A. Platonova¹, E. G. Medvedeva¹, Yu. V. Gaidamaka^{1,2},
S. Ya. Shorgin²

¹Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya
St, Moscow, 117198, Russian Federation

²Federal Research Center "Computer Science and Control" of the Russian Academy
of Sciences (FRC CSC RAS), 44-2 Vavilov St, Moscow, 119333, Russian Federation

bobrikova-ev@rudn.ru, platonova-aa@rudn.ru, medvedeva-eg@rudn.ru,
gaydamaka-yuv@rudn.ru, sshorgin@ipiran.ru

Abstract

The paper proposes a method for assigning a modulation-code scheme by a base station scheduler on unmanned aerial vehicle, based on predicting the value of the signal-to-interference ratio on the equipment of a mobile user at the next time slot from a sequence of known values of this ratio in the past. Prediction is performed using machine learning; for this, a single-layer neural network was built and applied to solve a multi-parameter optimization problem using the stochastic gradient method. The trained neural network for the predicted value of the signal / interference ratio allows the scheduler to select the modulation-code scheme correctly, thereby ensuring the level of quality of data transmission in the radio channel required for the provision of the service.

Keywords: SINR, machine learning, neural network

1. Introduction

One of the tasks that must be solved by the unmanned aerial vehicle (UAV) base station (BS) scheduler of LTE and LTE-A generation 5G wireless communication networks is the task of selecting a Modulation Coding Scheme (MCS) when transmitting data in a radio channel between the base station / access point and the user. The correct assignment of the MCS allows avoiding repeated data transmission, which leads to inefficient use of the resources of the receiving-transmitting equipment and reduces the quality indicators of the service to the user, in particular, increases the delay in providing the service. The task of SINR prediction becomes especially

This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipients A.P., Section 2). The reported study was funded by RSF, project No.20-07-01064 (recipient E.M., Sections 3).

important due to the limited capacity of the UAV battery, when repeated requests due to incorrect MCS assignment at the next cycle reduce the UAV lifetime. In the case of stationary users, the task of selecting the MCS is not difficult, since the MCS is determined by the Channel Quality Indicator (CQI) value, which the base station scheduler periodically receives from the user equipment [1]. CQI depends on the Signal to Interference and Noise Ratio (SINR) value on the user's equipment, which is affected by the radio access technology, the distance between transceivers, the radio signal propagation environment, including the presence of obstacles and signal blockers, the power of interfering transmitters, etc. When the mobile user moves, the user's geolocation, the distance to the BS and the parameters of the radio channel change, after which the receiver fixes and sends the base station a new CQI value, according to which the BS scheduler must adjust the MCS for transmission from the BS. Thus, the problem arises of predicting the CQI value of a mobile user. This should allow the scheduler to avoid the inefficient default MCS selection scheme. When, upon receiving, the CQI is lower than that required to provide the service, the scheduler traverses successively adjacent MCSs until it receives a CQI value from the user equipment that is acceptable for data transmission with adequate quality. The paper proposes an approach to solving this problem using the apparatus of neural networks [2].

2. System model and mathematical problem statement

One of the most popular machine learning methods is the apparatus of neural network. The simplest single-layer neural network - a single-layer perceptron [3] - will be used in the paper when predicting the MCS value of a mobile user based on known MCS values in the past. A system model is investigated, consisting of one cell of a wireless communication network, divided into square cells, one of which houses a BS, M is the number of square cells. We consider the user's movement model to be known - the data set of movement trajectories is either specified from observations, or generated using an analytical movement model, for example, in the form of a Grid Random Walk model in a discrete time. The user's trajectory is the final sequence of cell indicators in which the user was on the corresponding time step.

We believe that the cell indicator uniquely determines the distance between the BS and the user equipment, then the user's trajectory can be set by a sequence of MCS values, or by a sequence of CQI values for user equipment on adjacent time steps, or even by the Signal to Interference and Noise Ratio (SINR), which uniquely determines the CQI value. In machine learning terminology, we can say that the approach using a neural network allows for a given training sample - the user's movement trajectory - to predict the answer, i.e. the cell indicator at the next time step. As a mathematical model for such a prediction, a single-layer neural network

was chosen, which, as a result of solving a multi-parameter optimization problem based on the data set given at the input, is trained to output an approximate answer that meets the specified criteria.

In the course of the research, it was found that greater accuracy of predictions is achieved when choosing as precedents of the training sample not the numbers of cells, but the values of the signal / interference ratio, which is illustrated when constructing a mathematical model, as well as in a numerical example in the next section of the paper.

Let the object \mathbf{x}_i be a set of n sequential timeslots, l is the number of objects, $i = 1, \dots, l$. Let's introduce the attributes of the object \mathbf{x}_i : $f_1(\mathbf{x}_i), \dots, f_n(\mathbf{x}_i)$ are known SINR values corresponding to n timeslots of the object \mathbf{x}_i , i.e. n is the number of features. Let's denote $\mathbf{f}(\mathbf{x}_i) = (f_1(\mathbf{x}_i), \dots, f_n(\mathbf{x}_i))$. Let's consider the known SINR value as the answer y_i in the next fifth timeslot. Note that the number of responses is equal to the number of objects, the pair object and the response (\mathbf{x}_i, y_i) is called a precedent, the set of pairs $X^l = (\mathbf{x}_i, y_i)_{i=1}^l$ - training set.

As an approximating algorithm, we choose the linear model

$$a(\mathbf{x}, \mathbf{w}) = \langle \mathbf{w}, \mathbf{f}(\mathbf{x}) \rangle = \sum_{j=1}^n w_j f_j(\mathbf{x}), \quad (1)$$

where \mathbf{x} is an object, n is the number of features, w_j are unknown feature weights, $j = 1, \dots, n$.

The process of selecting the optimal parameter \mathbf{w} from the training set X^l is called the training of the a algorithm. The optimal parameter \mathbf{w} of the model is the parameter that provides the minimum value to the quality functional $Q(a, X^l)$ of the a algorithm on the sample X^l , that is

$$Q(a, X^l) = \frac{1}{l} \sum_{i=1}^l \mathcal{L}(a, y_i) \rightarrow \min_{\mathbf{w}}. \quad (2)$$

Here the function

$$\mathcal{L}(a, y_i) = (a(\mathbf{x}_i, \mathbf{w}) - y_i)^2 \quad (3)$$

is called the loss function and reflects the accuracy of the approximation on the object \mathbf{x}_i .

For the numerical optimization of functional (2), the work uses the stochastic gradient method [2], [4].

The described process of selecting the optimal parameter \mathbf{w} based on the training set $X^l = (\mathbf{x}_i, y_i)_{i=1}^l$ is the operation of the simplest single-layer neural network. For the object \mathbf{x}_i , the result of applying the linear function $a(\mathbf{x}_i, \mathbf{w})$ is called the

predicted SINR value on the object \mathbf{x}_i or simply the prediction on the object \mathbf{x}_i . Note that $a(\mathbf{x}_i, \mathbf{w})$ is an approximation of the known value y_i of the unknown function $y = y(\mathbf{x})$ on object \mathbf{x}_i .

3. An example of using the neural network apparatus for predicting the signal-to-interference ratio

An illustration of a neural network's operation for predicting the signal / interference ratio in a cell of an LTE network was carried out for an example of sampling trajectories of user's movement. For the BS coverage area of $M = 112320$ cells, the Monte Carlo method generated 100 trajectories according to the Grid Random Walk motion model [5] in the form of cell indicator's sequences of different lengths. Further, for the LTE technology, according to [1, 5], from the trajectories, the corresponding sequences of values of the signal / interference ratio were obtained, and from them - "fives", i.e. precedents which constituted a training sample of length $l = 38398$. The precedents that make up the training set are "object-response" pairs, where the object is a set of four sequential timeslots, a vector of features of length $n = 4$ is given for each object, and the features and responses are the values of the signal / interference ratio.

For example, a sequence of SINR values

$$\dots 8.26 \rightarrow 8.26 \rightarrow 8.33 \rightarrow 8.41 \rightarrow 8.48 \rightarrow 8.56 \rightarrow 8.64 \rightarrow 8.72 \dots$$

gives a training set with the following "object - response" pairs:

$$(8.26, 8.26, 8.33, 8.41) \rightarrow 8.48; \quad (8.26, 8.33, 8.41, 8.48) \rightarrow 8.56; \\ (8.33, 8.41, 8.48, 8.56) \rightarrow 8.64; \quad (8.41, 8.48, 8.56, 8.64) \rightarrow 8.72 \text{ etc.}$$

The initial values of the weights of features for the linear model (1) are chosen $\mathbf{w} = (0.25, 0.25, 0.25, 0.25)$, the gradient step (learning rate) $h = 10^{-4}$ [6].

Let us consider the prediction results from information about the SINR values of the user in 4 consecutive timeslots of the SINR value on the user's equipment in the next 5th timeslot using the constructed neural network.

Figure 1 shows the dotted line the predictions $a(\mathbf{x}_i, \mathbf{w})$, the solid line - answers y_i for the first 1000 objects of the training set. The criterion for stopping the training of a neural network can be stabilization of one of the following model parameters - weights, quality functional, accuracy. It took 25 epochs to train, and the maximum accuracy equal to 0.692 was achieved already at the 20th epoch (Fig. 2). Here, the epoch is one iteration of the stochastic gradient method for all objects of the training sample.

The coincidence of the predicted values and responses in Fig. 2 also confirm the prediction quality indicators: the quality functional (2) (Fig. 3) and the accuracy,

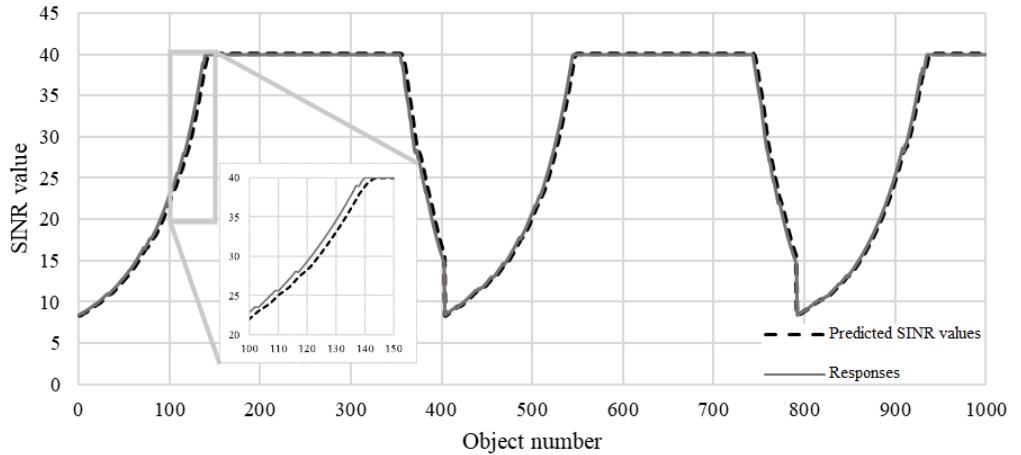


Fig. 1. Predicted SINR values and responses for the first 1000 objects in the training set

defined as the ratio of the number of coincident predicted values and responses to the total number of objects in the training sample, and we consider the case $\mathcal{L}(a, y_i) < 1$.

4. Conclusion

The proposed method for assigning a modulation-code scheme by the base station scheduler, based on predicting the value of the signal-to-interference ratio using the apparatus of neural networks, consists of two stages. At the first stage, for a given user movement model, a single-layer neural network for predicting the SINR value on the mobile user's equipment based on the known values of this ratio in the past is built and trained. At the second stage, according to the predicted value of the signal-to-interference ratio, the MCS is determined, which is required to transmit data to the user when providing a service with an appropriate level of quality. Note that the optimization problem, which is solved when training the network, is multivariable, while the parameters of the neural network significantly depend on the user movement model. In the future, to solve the problem of predicting the signal-to-interference ratio, it is planned to both develop supervised learning methods, for example, apply a multilayer neural network, and use the reinforcement learning method for prediction in the absence of a training sample.

The authors are grateful to the MSc students of the Applied Informatics and Probability Theory Department of the RUDN University E. M. Khairov and A. A. Mamonov for preparing a training sample for a numerical experiment.

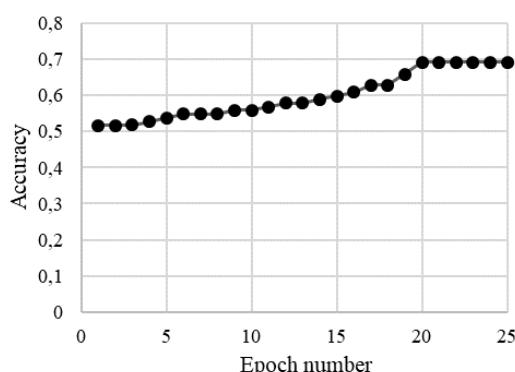


Fig. 2. Accuracy

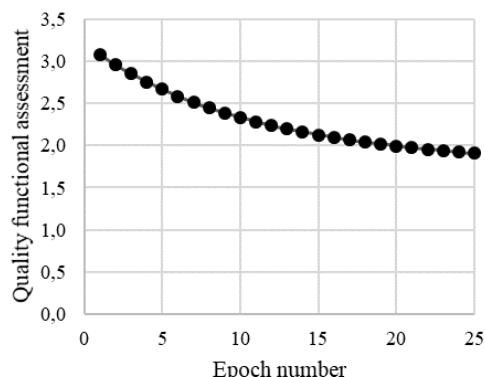


Fig. 3. Changing the assessment of the quality functional

REFERENCES

1. Ghosh A., Ratasuk R. Essentials of lte and lte-a // Cambridge University Press, 2011.
2. Vorontsov K. Matematicheskie metody obucheniya po precedentam (teoriya obucheniya mashin) [Mathematical teaching methods by precedents (machine learning theory)], 2011. <http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf>
3. Averkin A., Gaaze-Rapoport M. G., Pospelov D. A. Tolkovyj slovar' po iskusstvennomu intellektu [Dictionary of Artificial Intelligence], Radio i svyaz' [Radio and communication], 1992.
4. Trask A. W. Grokking deep learning // Simon and Schuster, 2019.
5. Gaidamaka Y., Samouylov K., Shorgin S. Method of modeling interference characteristics in heterogeneous fifth generation wireless networks with device-to-device communications // Inform. primen. 2017. V. 11. P. 2–9.
6. Bobrikova E., Platonova A., Yartseva I., Khairov E. K zadache predskazaniya sinr v besprovodnoj seti s podvizhnymi pol'zovatelyami s pomoshch'yu apparatanejronnyh setej [To the problem of predicting sinr in a wireless network with mobile users using a neural network apparatus], Informatsionno-telekommunikatsionnye tekhnologii i matematicheskoe modelirovanie vysokotekhnologichnykh sistem [Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems 2021]. 2021. P. 30–32.

UDC: 004.89

State Observer System Based on K-Means Clustering Machine Learning Model for Cyber-Security of Industrial Network

Artur Sagdatullin¹

¹Kazan National Research Technical University named after A.N. Tupolev,
Kazan, K. Marx 10 Str., Russia
saturn-s5@mail.ru

Abstract

Machine learning is a good reasoning automated mechanism to help human intelligence to detect potential risks. There are the following ways used techniques could be implemented: spam detection based on the texts analysis, anomalies detection and data prediction. In this paper it'll be discussed the text analysis algorithm based on the clustering machine learning technique for detecting potential area of threaten messages in the modeled industrial system. It is implemented a discrete-time observer algorithm of machine learning model for cybersecurity of industrial network.

Keywords: Machine learning, cybersecurity, discrete-time observer, industrial network

1. Introduction

In the last decades, there has been significant interest in cybersecurity as today's world of interconnected systems potentially a targets for emerging threats. Industrial applications and networks are automated by internet technologies and internet of things devices. In this environment, it's hard for assessing of hidden threats. Also, last investments of researchers of all the world in computer science spawn the vast area of artificial intelligence systems such as machine learning and learning algorithms. Application based on machine learning are used in different areas of human being. One of the prospective field of implementing such a technologies is cybersecurity. Using specially intended for cybersecurity machine learning models researchers could detect potential threat and malware. There are many definitions for cybersecurity field relates to, but in this paper it comprises not only information about inner network and devises, but also outer network security.

Machine learning is a good reasoning automated mechanism to help human intelligence to detect potential risks. There are the following ways used techniques could be implemented: spam detection based on the texts analysis, anomalies detection and data prediction. In this paper it'll be discussed the text analysis algorithm based on the clustering machine learning technique for detecting potential area of threaten messages in the modeled industrial system.

2. State observer architecture for cybersecurity of industrial network

2.1. Discrete-time system mathematical definition. The significant coupling of objects and components of the system can be a considerable difficulty, both in software development and in the implementation of automated systems. For a continuous linear time-invariant system, equations are following:

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{Bu} \\ \mathbf{y} &= \mathbf{Cx}\end{aligned}\tag{1}$$

Unknown parameter $X(t)$ as a random variable with its own expectation $E(x)$ and distribution could be defined as:

$$\begin{aligned}\mathbf{x}[k];x &\geq 0 \\ \mathbf{x}[k] &= \mathbf{x}[k - 1] + \mathbf{z}[k]\end{aligned}\tag{2}$$

where $\mathbf{z}[k]$ is a noise process, supposed to be $\sim N(0, \sigma_{nn}^2)$. Consider a LTI system with discrete time:

$$\begin{aligned}\mathbf{x}(k) &= \mathbf{A}(k - 1)\mathbf{x}(k - 1) + \mathbf{f}(k) \\ \mathbf{z}(k) &= \mathbf{C}(k)\mathbf{x}(k) + \mathbf{w}(k)\end{aligned}\tag{3}$$

where $\mathbf{x}(k)$ represents a state variables vector, $\mathbf{A}(k-1)$ is the transition matrix, $\mathbf{f}(k)$ is a noise vector inherited to state space model, $\mathbf{z}(k)$ is the observations vector, $\mathbf{C}(k)$ is a state matrix related to $\mathbf{x}(k)$.

For sampling interval of T linear discrete-time system equations are following:

$$\begin{aligned}\dot{\mathbf{x}}(k - 1) &= \mathbf{Lx}(k) + \mathbf{Hu}(k) \\ \mathbf{y}(k) &= \mathbf{Cx}(k) \\ \mathbf{Lx} &= e^{\mathbf{AT}} \\ \mathbf{H} &= e^{\mathbf{AT}} \int_0^T e^{-\mathbf{A}\tau} d\tau\end{aligned}\tag{4}$$

where \mathbf{u} is the control vector, \mathbf{x} is the state vector, \mathbf{L} is the unobservable perturbations, \mathbf{H} is the function of disturbing influences on the control object.

This state observer model may control incoming data for the LTI system. It allows predicting the system's state based on the $(k - 1)$ time-step measurement.

2.2. State system observer control scheme. The representation of state vectors describes the characteristics of the control object for closed-loop systems with feedback, the tasks of which are to find the optimal location of the roots of the system equation by adjusting its parameters. Digital representation is the direct form of implementing state observers as software modules. Thus, a discrete description will be implemented for digital representation. Therefore, to describe such systems, it is required to represent the discrete state of the system (see Fig. 1).

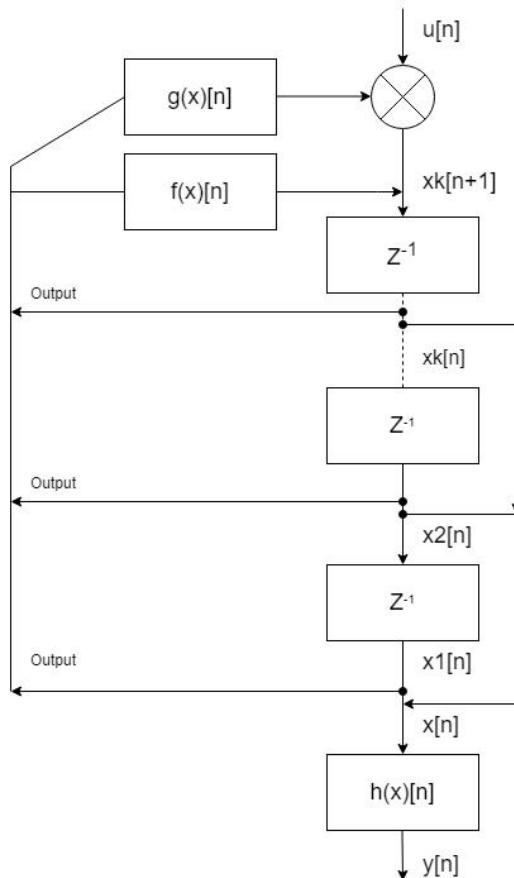


Fig. 1. State system observer control scheme

The State system observer control scheme solves the problem of monitoring parameters (A, B, C, D) of the industrial system at any given time. The system's state describes an estimate of the measurement of input and output. This expression shows that under known initial conditions, the state estimate is approximately equal to the actual value of the output signal. This fact gives us an observability definition concerning the designed machine learning model to process input text messages as parameter values.

3. Methodology of machine learning algorithm design

Machine learning algorithms comprise linear and logistic regression, polynomial regression, decision trees, support vector machines, random forests, etc. All these concepts are the tasks of supervised and unsupervised learning. Classification and clustering are the methods at the heart of the proposed system. The Clustering aims to distinguish n clusters in $[k - 1]$ observations and distribute the rest of estimates value points in the nearest mean cluster. An algorithm of the k-means clustering consists of a series of steps to determine the distance between points and clusters:

Algorithm 1 An algorithm according to the methodology

```

Require:  $n \geq 0$                                      ▷ This is the number of clusters
Ensure:  $k \geq 0$                                      ▷ Centroid data points
while  $i \neq 0$  do
    if  $X_i$  then
         $L(x_1, x_2) \leftarrow \sum_i^{[x_1] \in \mathcal{R}^N} (X_1[i], X_2[i])^2$  ▷ Categorize each point to its closest
        centroid
         $k \leftarrow L(x_1, x_2)$ 
    else
         $k \leftarrow k$                                          ▷ Compute the new centroid
         $i \leftarrow i - 1$ 
    end if
end while

```

The algorithm of the clustering problem states that by a given vector of parameters, it could be expected centroid distance to individual data points. Then, the mean of the vector will be recomputed for a new centroid finding.

K-means clustering algorithm performance is shown in Fig. 3. Training takes about 0.17 seconds for the two cluster analysis and about 0.23 seconds for three cluster centroid problems with samples of about 5000.

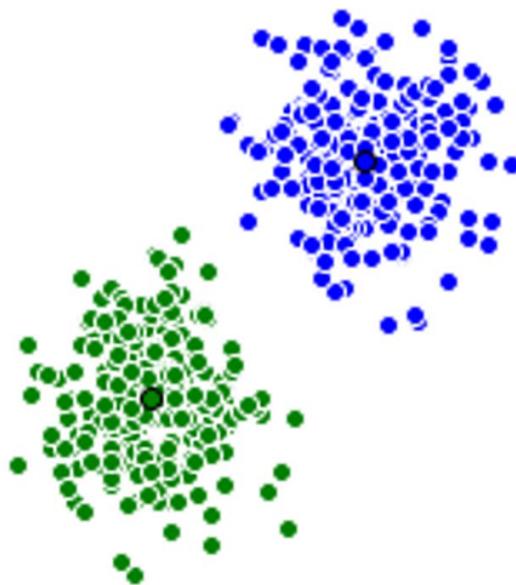


Fig. 2. Two cluster analysis problem



Fig. 3. Three cluster analysis problem

4. Experiment

For the experiment it was taking training data about potentially threaten and spam messages inside industrial network. All data results presented in figures (**Fig. 2** and **Fig. 3**).

5. Conclusion

Machine learning is a crucial factor for future cybersecurity and information and communication networks. This paper will discuss the text analysis algorithm based on the clustering machine learning technique for detecting potential areas of spam messages in the modeled industrial system. A discrete-time observer algorithm of the machine learning model for cybersecurity of industrial networks is implemented.

REFERENCES

1. Katare D, El-Sharkawy M. Embedded system enabled vehicle collision detection: an ANN classifier. In: 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC); 2019. p. 0284–0289.
2. Chen Y. , M. Khandaker, Z. Wang, in Proceedings of the 2017 ACM on AsiaConference on Computer and Communications Security. ASIA CCS '17. Pinpointing vulnerabilities (ACM, New York, 2017), pp. 334–345
3. Ishida C. , Y. Arakawa, I. Sasase, K. Takemori, in Proceedings of PACRIM. 2005IEEE Pacific Rim Conference on Communications, Computers and signal-Processing, August 24-26. Forecast techniques for predicting increase ordecrease of attacks using bayesian inference (IEEE, Victoria, 2005), pp. 450–453
4. Z. Li, D. Zou, S. Xu, X. Ou, H. Jin, S. Wang, Z. Deng, Y. Zhong, in 25th AnnualNetwork and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018. Vuldeepecker: A deep learning-basedsystem for vulnerability detection (Internet Society, San Diego, 2018)
5. Thennakoon, Anuruddha, et al. Real-time credit card fraud detection using machine learning. In: 2019 9th international conference on cloud computing, data science & engineering (Confluence). IEEE; 2019.

UDC: 519.218

Reliability Analysis of a k -out-of- n System in Case of Full Repair After Its Failure

Ivanova N.M.^{1,2}

¹Peoples' Friendship University of Russia (RUDN University),
6 Miklukho-Maklaya St., Moscow, 117198, Russian Federation

²V.A.Trapeznikov Institute of Control Sciences of Russian Academy of Sciences,
65 Profsoyuznaya street, Moscow, 117997, Russia

nm_ivanova@bk.ru

Abstract

The paper focuses on the investigation of a k -out-of- n system in case of full repair after its failure. The reliability analysis, including the calculation of time-dependent and steady-state system characteristics, is considered. It is assumed that components' lifetime is exponentially distributed, and the repair time has an arbitrary distribution. The closed-form representation of system steady-state probabilities in terms of the Laplace transform of components' repair time is shown.

Keywords: k -out-of- n system, reliability analysis, markovization method, steady state probabilities, sensitivity analysis, rare failures

1. Introduction

Nowadays, the issue of improving system reliability becomes more and more significant and relevant in connection with the complex mechanization and automation of processes occurring in numerous areas of human activity. The importance of this problem is due to the fact that the unsatisfactory reliability of the system generates large costs for its maintenance and, in some cases, can lead to serious consequences.

There are many methods aimed at increasing reliability. Among them, the redundancy technique is the most popular method. As an example of redundancy, the model of a k -out-of- n system can be considered. Such a system consists of n components and fails when at least k of them fail. This type of system is commonly referred to as a k -out-of- n : F system. So, this definition is implied throughout the

This publication has been supported by the RUDN University Scientific Projects Grant System, project № P02 IF, theme No. 021930-2-000 (review and analytic results).

The publication has been partially funded by RFBR according to the research project No. 20-01-00575A (problem setting) and RSF project No. 22-49-02023 (formal analysis, validation).

paper without the symbol “ F ”. Due to the wide applications of such systems in many spheres of human activity (engineering, telecommunication, medicine, management), many papers are devoted to their investigation. There is an extensive literature on the study of such systems (for example, see [1] and the bibliography within).

The k -out-of- n systems’ investigation has many publications in which these systems are considered under various assumptions about the shape of life and repair time distributions [2], the dependency and several types of failures [3], the repair type and its availability, arrangements of components, load sharing [4] and others.

In addition, in the context of reliability analysis, it is also of interest to analyze the sensitivity of their objective characteristics to the shapes of input distributions. For such a purpose, various analytical methods on the basis of multidimensional Markov processes as well as simulation ones [5] are used.

In order to continue research in the reliability field, the current paper considers a k -out-of- n system using one of the markovization method consists in the introduction of supplementary variables [6]. This method allows describing the system’s behavior by a two-dimensional Markov process that enables one to find analytical expressions for the system’s steady state probabilities (s.s.p.’s) that is the purpose of this work.

2. Notations and Assumptions

Consider a repairable k -out-of- n ($k < n$) system. A repairable system is one that is repaired not only after a component failure (partial repair), but also after the failure of the entire system (full repair). According to the possibility of the system’s restoration and its components, suppose that there is one unit for repair procedure, and after the restoration of all failed components (full failure), the system becomes as new. Introduce some assumptions about the shape of components life and repair time distributions. Suppose that

- the lifetimes of system components are exponentially distributed with parameter α and mean time $a = \alpha^{-1}$;
- the repair times for any failed components (the case of partial repair) are independent identically distributed (i.i.d.) random variables (r.v.’s) B_i ($i = 1, 2, \dots$) with common cumulative distribution function (c.d.f.) $B(x) = \mathbf{P}\{B_i \leq x\}$ which is absolute continuous with its probability density function (p.d.f.) $b(x)$;
- the repair times for failed system (the case of full repair) are also i.i.d. r.v.’s F_i ($i = 1, 2, \dots$) with corresponding c.d.f. $F(x) = \mathbf{P}\{F_i \leq x\}$, its p.d.f. is $f(x)$;

- the instantaneous repairs are impossible, their mean times are finite,

$$B(0) = F(0) = 0, \quad b = \int_0^\infty (1 - B(x))dx < \infty, \quad f = \int_0^\infty (1 - F(x))dx < \infty;$$

- corresponding Laplace transforms (LTs) of the r.v.'s B and F are

$$\tilde{b}(s) = \int_0^\infty e^{-st} dB(t), \quad \tilde{f}(s) = \int_0^\infty e^{-st} dF(t).$$

Let i is the number of failed components out of n , at that, according to the repair rule, one of such components is repaired. Thus, $(n - i)$ is the number of working components. Such a states' space can be represented as $E = \{0, 1, 2, \dots, k - 1, k\}$. Here the state 0 shows that all the components are working, no one is repaired. The state k means the full system failure and its repair with r.v. F , after which the system becomes as a new one.

To perform reliability analysis, submit a random process $J = \{J(t), t \geq 0\}$ on a space set E as a description of the system behavior,

$$J(t) = j, \quad j \in E, \quad \text{if the system is in state } j \text{ at time } t.$$

Suppose that $J(0) = 0$. The paper is devoted to time-dependent system state probabilities (t.d.s.s.p.'s) $\pi_j(t)$ and s.s.p.'s π_j calculation,

$$\pi_j(t) = \mathbf{P}\{J(t) = j\}, \quad \pi_j = \lim_{t \rightarrow \infty} \mathbf{P}\{J(t) = j\}, \quad j \in E,$$

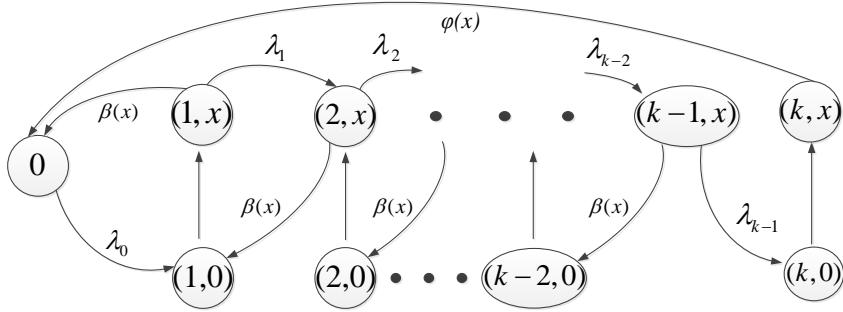
as well as properties of their asymptotic sensitivity to the shapes of system components' repair time distribution.

3. Main Results

Present analytical results of the s.s.p.'s calculation of a k -out-of- n system with the help of the method of supplementary variables [6]. In the case under consideration, as a supplementary variable, the elapsed repair time of the failed component is used. Thus, denote by

$$Z(t) = \{J(t), X(t)\}_{t \geq 0}$$

a two-dimensional Markov process with extended states' space $\bar{E} = \{0, (i, x) \mid i = \overline{1, k}\}$. In this notation $J(t)$ is defined as above, and $X(t)$ means the elapsed repair time of the failed component or the whole system. Figure 1 represents the states' transition graph of the process $Z(t)$.

Fig. 1. Transition graph of the process $Z(t)$

Here the system failure intensity, when i components out of n fail, is denoted by $\lambda_i = (n - i)\alpha$, ($i = \overline{0, k-1}$). Moreover, assuming elapsed repair time x , the conditional intensities of partial and full repair are, respectively,

$$\beta(x) = \frac{b(x)}{1 - B(x)}, \quad \phi(x) = \frac{f(x)}{1 - F(x)}.$$

Denote by

- $\pi_0(t) = \mathbf{P}\{J(t) = 0\}$ – the probability of a working state of all system components at time t ;
- $\pi_i(t; x) = \mathbf{P}\{J(t) = i; x < X(t) \leq x + dx\}$ – the joint probability that at time t there are i failed components, among which one is repaired with the elapsed repair time in the interval x and $x + dx$, $i = \overline{1, k}$.

From the graph 1 as well as by comparing the process $Z(t)$ in the closed interval t and $t + \delta$, the following Kolmogorov forward system of partial differential equations for the t.d.s.s.p.'s calculation are obtained.

Theorem 1. T.d.s.s.p.'s of the process $Z(t)$ for $k > 2$ are followed from the system of Kolmogorov forward partial differential equations (1),

$$\begin{aligned} \frac{d}{dt}\pi_0(t) &= -\lambda_0\pi_0(t) + \int_0^t \beta(x)\pi_1(t, x)dx + \int_0^t \phi(x)\pi_k(t, x)dx, \\ \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right)\pi_1(t, x) &= (\lambda_1 + \beta(x))\pi_1(t, x), \\ \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right)\pi_i(t, x) &= -(\lambda_i + \beta(x))\pi_i(t, x) + \lambda_{i-1}\pi_{i-1}(t, x), \quad i = \overline{2, k-1}, \\ \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right)\pi_k(t, x) &= -\phi(x)\pi_k(t, x), \end{aligned} \quad (1)$$

with the initial (2)

$$\pi_0(0) = 1, \quad \pi_i(0; x) = 0, \quad i = \overline{1, 3}, \quad \forall x \geq 0, \quad (2)$$

and boundary (3) conditions

$$\begin{aligned} \pi_1(t, 0) &= \lambda_0 \pi_0(t) + \int_0^t \beta(x) \pi_2(t, x) dx, \\ \pi_i(t, 0) &= \int_0^t \beta(x) \pi_{i+1}(t, x) dx, \quad i = \overline{2, k-2}, \\ \pi_{k-1}(t, 0) &= 0, \\ \pi_k(t, 0) &= \lambda_{k-1} \int_0^\infty \pi_{k-1}(t, x) dx. \end{aligned} \quad (3)$$

A further step in reliability analysis is calculation of s.s.p.'s. For this, there are at least two possibilities. From the first one, s.s.p.'s can be obtained by passing to $\pi_i = \lim_{s \rightarrow 0} \tilde{\pi}_i(s)$ from t.d.s.s.p.'s. In this paper, another way is applied with the following reasoning.

The process $Z(t)$ is a Harris one, so according to the Harris Markov processes theory, it has a stationary regime, and, therefore, for $t \rightarrow \infty$ its t.d.s.s.p.'s tends to corresponding s.s.p.'s. It means that the process $Z(t)$ has a stationary distribution for which a system of differential equations (balance equations) holds with corresponding transition in the systems (1)-(3). Such a transition and the application of the method of constants variation give the following result.

Theorem 2. The s.s.p.'s of the process $Z(t)$ for $k > 2$ have the form

$$\begin{aligned} \pi_0 &= \lambda_0^{-1} \left[C_1 \left(1 + \frac{\lambda_1}{\lambda_1 - \lambda_2} \tilde{b}(\lambda_1) \right) - C_2 \tilde{b}(\lambda_2) \right], \\ \pi_1 &= C_1 \frac{1 - \tilde{b}(\lambda_1)}{\lambda_1}, \\ \pi_i &= C_i \frac{1 - \tilde{b}(\lambda_i)}{\lambda_i} + S(i-1) \frac{1 - \tilde{b}(\lambda_{i-1})}{\lambda_{i-1}}, \quad i = \overline{2, k-1}, \\ \pi_k &= C_k \cdot f, \end{aligned}$$

where

$$\begin{aligned} C_i &= C_{i+1}\tilde{b}(\lambda_{i+1}) + S(i)\tilde{b}(\lambda_i) - S(i-1), \quad i = \overline{2, k-2}, \\ C_{k-1} &= -S(k-2), \\ C_k &= \lambda_{k-1}C_{k-1} \left(\frac{1 - \tilde{b}(\lambda_{k-1})}{\lambda_{k-1}} - \frac{1 - \tilde{b}(\lambda_{k-2})}{\lambda_{k-2}} \right), \\ S(i) &= \sum_{j=1}^i (-1)^{i-j+1} \left(\prod_{m=j}^i \frac{\lambda_m}{\lambda_j - \lambda_{m+1}} \right) C_j, \end{aligned}$$

and C_1 is calculated from the normalization condition $\sum_{i \in E} \pi_i = 1$.

Example 1. For example, present the case of a 3-out-of-6 system. From the Theorem 2 we get

$$\begin{aligned} \pi_1 &= \frac{6}{5} \cdot \frac{1 - \tilde{b}(5\alpha)}{1 + 5\tilde{b}(5\alpha) - 5\tilde{b}(4\alpha)} \pi_0, & \pi_2 &= \frac{3}{2} \cdot \frac{1 + 4\tilde{b}(5\alpha) - 5\tilde{b}(4\alpha)}{1 + 5\tilde{b}(5\alpha) - 5\tilde{b}(4\alpha)} \pi_0, \quad (4) \\ \pi_3 &= \frac{6\alpha f(1 + 4\tilde{b}(5\alpha) - 5\tilde{b}(4\alpha))}{1 + 5\tilde{b}(5\alpha) - 5\tilde{b}(4\alpha)} \pi_0, & \pi_0 &= 1 - \sum_{1 \leq i \leq 3} \pi_i, \end{aligned}$$

and the coefficient of availability

$$K_{av} = 1 - \pi_3 = \frac{37 + 58\tilde{b}(5\alpha) - 75\tilde{b}(4\alpha)}{60\alpha f(1 + 4\tilde{b}(5\alpha) - 5\tilde{b}(4\alpha)) + 37 + 58\tilde{b}(5\alpha) - 75\tilde{b}(4\alpha)}.$$

It should be noted that for exponential distribution of partial repair time $B(x) = 1 - e^{-\beta x}$ with LT $\tilde{b}(\alpha) = \beta(\alpha + \beta)^{-1}$ and mean time $b = \beta^{-1}$, as well as distribution-independent mean full repair time f , the probabilities (4) match with those, obtained with a simple birth and death process,

$$\begin{aligned} \pi_0 &= \frac{20\alpha^2 + 4\alpha\beta + \beta^2}{120\alpha^3 f + 74\alpha^2 + 10\alpha\beta + \beta^2}, & \pi_1 &= \frac{6\alpha(4\alpha + \beta)}{20\alpha^2 + 4\alpha\beta + \beta^2} \pi_0, \\ \pi_2 &= \frac{30\alpha^2}{20\alpha^2 + 4\alpha\beta + \beta^2} \pi_0, & \pi_3 &= \frac{120\alpha^3 f}{20\alpha^2 + 4\alpha\beta + \beta^2} \pi_0. \end{aligned}$$

4. Conclusion

An explicit form of the stationary probabilities of a k -out-of- n system is presented in terms of the Laplace transform of the components' partial repair time. The expressions show their dependence on the shape of partial repair time distribution

and independence on the shape of full repair time distribution. The results obtained are verified using the Markov case.

Further investigations will continue the reliability analysis of a k -out-of- n system. The asymptotic behavior of the system under rare failures will be considered. The sensitivity of system reliability characteristics to the shape of the component's repair time distribution and corresponding coefficient of variation will be investigated by numerical analysis.

REFERENCES

1. Zuo, M.J., Tian, Z. k -out-of- n Systems, Wiley Encyclopedia of Operations Research and Management Science, edited by James J. Cochran, 2010
2. Rykov, V.V, Ivanova, N.M. Reliability and sensitivity analysis of a repairable k -out-of- n :F system with general life- and repair times distributions. // Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference. Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio. Singapore: Research Publishing Services, doi: 10.3850/978-981-14-8593-0_5750-cd, 2020
3. Linmin Hu, Sijia Liu, Rui Peng & Zhaocai Liu. Reliability and sensitivity analysis of a repairable k -out-of- n :G system with two failure modes and re-trial feature // Communications in Statistics – Theory and Methods, doi: 10.1080/03610926.2020.1788083, 2020
4. Amari, S., Bergman, R. Reliability analysis of k-out-of-n load-sharing systems // Annual Reliability and Maintainability Symposium, 440-445, doi: 10.1109/RAMS.2008.4925836, 2008
5. Ivanova N. Modeling and Simulation of Reliability Function of a k-out-of-n:F System // Communications in Computer and Information Science, V.1337. doi: 10.1007/978-3-030-66242-4_22, 2020
6. Kalashnikov, V.V. Method of Supplementary Variables. In: Mathematical Methods in Queuing Theory. Mathematics and Its Applications, vol 271. Springer, Dordrecht. doi: 10.1007/978-94-017-2197-4_10, 1994

UDC: 519.23

Closed Stochastic Network of the Needham-Schroeder model for Oil Pipeline Data Transmission

A.A. Shukmanova¹, A.S. Yermakov², T.T. Paltashev³, A.K. Mamyrova⁴

¹Kazakh National Technical University, Almaty, Kazakhstan

²Caspian University, Almaty, Kazakhstan

³ITMO University, Saint Petersburg, Russia

⁴Turan University, Almaty, Kazakhstan

ermakov_as@mail.ru, mamyrova_ak09@mail.ru

Abstract

The article presents a model of a closed stochastic pipeline network. The calculation algorithm based on the Needham-Schroeder protocol is derived. The calculation method based on two-phase systems such as AISO, SIAO is derived.

Keywords: model, pipeline, control and Needham-Schroeder algorithm, closed network and parameterization, AISO methods, SIAO

1. Pipeline operation architecture

The operation of a pipeline always involves a combination of the need for data transmission, the efficiency of the pipeline system, and the demand for inventory in tank farms or at delivery points.[1] Therefore, correct and efficient transportation planning is essential for the successful and reliable operation of each pipeline. Using the Message Planner and an easy-to-use tool for preparing transportation plans can assist in much faster planning and avoidance of pipeline operation collisions. The plan that used to take days to prepare now takes minutes. Moreover, the transportation plan prepared by the Message Planner can be transferred to the Message Tracking System's scheduled message database.

The Emergency Shutdown System as a safety instrumented system is designed to prevent and mitigate the consequences of emergency situations in the pipeline operation. Its failure or absence poses an unacceptable risk to the pipeline, people, and the environment. When a trigger condition is detected, the Emergency Shutdown System immediately and safely stops (isolates) the pipeline.

Fast and reliable communications form the backbone of the pipeline management system and are critical in achieving the required quality parameters.

The key points of pipeline supervision are continuous monitoring pipelines in stationary and transient states. This requires tracking out the maximum working pressure, detecting leaks, and monitoring parameters on a real-time basis.

The Message Tracking System is designed to ensure accurate delivery performance along with optimal pipeline operation. It also provides other parts of the system with accurate product data. This means that leak detection and pressure tracking systems operate with accurate product data.

The functionality of the hydraulic profile allows keeping the pipelines in a "healthy" working condition. This system provides operators with online information about the actual hydraulic profile of each pipeline section and generates an alarm whenever the pressure in any section of the pipeline

2. Protocol "Alice → Bob"

Alice initiates the protocol. First of all she generates a random session key and encrypts it with a shared key known to her and Trent. Then she sends Trent a ciphertext indicating her and Bob's names. Upon receiving Alice's request to deliver the session key Trent finds in his database the two shared long-term keys mentioned in Alice's request. After that Trent decrypts the resulting text using Alice's key, encrypts it again with Bob's key, and sends the result to Bob. Finally, having received and decrypted the delivered session key, Bob confirms its receipt by sending Alice a message encrypted with this key. In this protocol Alice is the **initiator** and Bob is the **responder**.

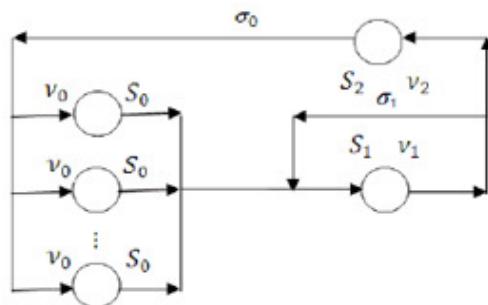


Fig. 1. Structure diagram of the Alice and Bob key exchange algorithm

A closed one-time password protection network based on the Needham-Schroeder algorithm [3]. It consists of systems S_0 (one or a group of M_0) users; S_1 is an intermediary or authorized person; and S_2 is the message destination. The user logs

in from the S_0 system with the transfer of the data byte to the destination in the S_2 system.

The entry is carried out through an intermediary in the S_1 system, which encodes the received byte and transmits it via a closed channel to the addressee in the S_2 system. The coding time is ν_1 , the time of transmission and processing by the destination is ν_2 . After this time the user receives a response from the destination with probability σ_0 .

To calculate the network characteristics, the following parameters must be set:

1. The list of systems that make up the network (systems S_0 , S_1 , and S_2).
2. Average times of servicing requests by systems (ν_0 , ν_1 , and ν_2).
3. Transmission probability matrix

$$\mathbf{P} = \mathbf{P}_{ij} = \begin{vmatrix} p_{00} & p_{02} & \dots & p_{0j} \\ p_{10} & p_{11} & \dots & p_{1j} \\ \vdots & & & \\ p_{i1} & p_{i2} & \dots & p_{ij} \end{vmatrix}$$

4. The number of request M passed around in the network.

The given subnetwork covers a special case of the Needham-Schroeder algorithm which consists of several stages depending on the considered options for implementing the algorithms. In particular, the implementation of subnetworks depends on the organization of input/output and the way the processor accesses the memory as well as the implementation of network data exchange functions in the network including the presence of DRONs. In the presented scheme, the general line of calculations is given.

The calculation of the stochastic network characteristics begins with determining the probabilities of states according to the formula:

$$P(M_0, M_1, M_2) = \frac{M! \gamma_1^{M_1} \gamma_2^{M_2}}{(M_0 \sum_{A(M,n)} \frac{M! \gamma_1^{M_1} \gamma_2^{M_2}}{M_0!})} \quad (1)$$

Where $P(M_0, M_1, M_2)$ is the probability of distribution of the number of requests M in network systems with M_0 , M_1 , and M_2 requests in each system. The factors $\gamma_j^{M_j}$ define the number of calls to the j-th systems in the network. They are determined as follows:

$$\gamma_j = \frac{\alpha_j \nu_j}{\nu_0} \quad (2)$$

where α_j defines the average number of calls to the j-th system per one request service. The value $A(M,n)$ defines the number of sets, depending on the number of requests M_n in systems n:

$$A(M, n) = \frac{(M + n - 1)}{(M!(n - 1)} \quad (3)$$

The main indicator for online service is the user downtime rate while waiting for a response η_0 :

$$\eta_0 = \sum_{r=1}^M \frac{r}{M} P(M_i = r) \quad (4)$$

where $P(M_i = r)$ is defined as the cumulative probability that all requests are online and no response has been received. This probability is calculated according to the formula (5):

$$P(M_i = r) = \sum_{M_i=r} P(M_1, M_2) \quad (5)$$

for $r = \overline{0, M}$

The average dwell time of a request in the system is equal (only the request dwell time in systems S_1 and S_2 is taken into account) as well as the average waiting times in the input and output buffers. This time is defined by the value of T_s and is equal to:

$$T_s = \sum_{i=1}^2 \sigma_i P(M = \sigma_i) \quad (6)$$

The performance of the system W is calculated through the stream rate of the served requests:

$$W = \frac{1 - P(M_1 = 0)}{\nu_1} \quad (7)$$

3. Network parameterization

3.1. AISO parameterization. In general, the option under consideration reflects the beginning of the Needham-Schroeder algorithm while the more complete option will be developed later. Systems S_1 and S_2 are three-phase QS_s with mixed service times and are difficult to calculate.

In the first approximation, a method is proposed for the equivalent replacement of a pair of synchronous phases with one equivalent in terms of the average service time.

First, we consider the AISO model, which stands at the network entry (S1 system). In this system, its second and third phases - due to the synchronization of service and on the basis of the mean value theorem - are replaced by an equivalent phase with the service rate $\mu_2^* = 1/\nu_2^*$. Here $\nu_2^* = \nu_2 + \nu_3$ is determined at average times for $\nu_1 = 1/\mu_i$, $i = \overline{2,3}$. This model interprets asynchronous input with processing or output and is shown in Figure 2.

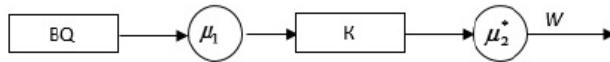


Fig. 2. Equivalent AISO model

For the analysis based on [4] the probabilities of the states $P_{ij}(n)$ of the first $i = \{1, \beta\}$ and the second $j = \{1, 0\}$ phases for $n = 0, 1, \dots, k$ requests in the buffer. Here i and j in states 1 define services in the appropriate phases. The state $i = \beta$ corresponds to the blocking of service in the first phase due to the buffer being busy, and the state $j = 0$ corresponds to waiting due to the absence of requests in the buffer. The Markov process for these states is represented by the final graph of states and transitions as shown in Figure 3.

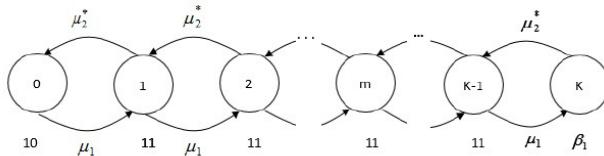


Fig. 3. Graph of states and transitions for AISO model

The graph nodes are marked by the number of requests in the buffer $m = 0, K$ and the states of the service phases i and j while the graph arcs correspond to the service rates. The system of linear algebraic balance equations for a stationary service mode, with notations $X_n = P_{ij}(n)$; $n = 1, 1, \dots, K$, is as follows(8) probabilities of

states where $\rho = \mu_1/\mu_2^*$. The determined probabilities are used to find the main characteristics:

- Overlap factor of input and processing or output

$$\left. \begin{array}{l} (\mu_1 + \mu_2^*)X_n = \mu_1 X_{n-1} + \mu_2^* X_{n+1}, n = 1, K-1 \\ \mu_1 X_{K-1} = \mu_2^* X_K \end{array} \right\} \quad (8)$$

$$\left. \begin{array}{l} X_n^* = \rho^n X_0, n = \overline{1, K} \\ X_0 = \sum_{q=0}^{K-1} \rho^q = (\rho^{k+1} - 1)/(\rho - 1) \end{array} \right\} \quad (9)$$

$$\epsilon_1 = \sum_{i=1}^{K-1} X_i = \rho(\rho^{k-1} - 1)/\rho(k+1) - 1, \quad (10)$$

- Loading factor of the input phase $\eta_1 = (1 - X_K)$,
- System throughput

$$\omega = \mu_1 \eta_1 = \mu_1 (\rho^K - 1)/(\rho^{k+1} - 1). \quad (11)$$

Based on equality (6), the load factors of the second and third phases of the model as shown on Figure 2

$$\eta_1 = \eta_1 \mu_1 / \mu_1, i = \overline{2, 3}. \quad (12)$$

3.2. SIAO parameterization. This system reflects the second stage of service according to the Needham-Schroeder algorithm [5,6,7]. It corresponds to the path with probability 0 that characterizes the end of service. The model, as mentioned above, is based on two-phase *QS*. Only an equivalent substitution is made for the first and second phases resulting in the framework with the replacement of the first and second phases. The load factor in this case is determined by the formula $\rho_1 = \mu_1^*/\mu_2$. See the procedure for formulas derivation above (1-11). In Table 1 for some combinations of values that satisfy the condition $\sum_{i=1}^3 \mu_i = 6$, the load factors n ρ of the equivalent two-phase models, relevant values of the maximum throughput

μ_1	μ_2	μ_3	ρ_{AISO}	ρ_{SIAO}	$\omega_{max}(AISO)$	$\omega_{max}(SIAO)$	ω_{SISO}
1	4	1	1,25	0,8	0,8	0,8	0,4444
1	2	3	0,43	0,22	1	0,67	0,5454
3	2	1	4,5	1,2	0,67	1	0,5454
2	2	2	2	0,5	1	1	0,6666

Table 1. Load factors for AISO and SIAO models

$\omega_{max} = \min\{\mu_1, \mu_2^*\}$ (AISO model), and $\omega_{max} = \min\{\mu_1^*, \mu_3\}$ (SIAO model) are given as well as the performance for the synchronous case calculated according to the formula (6).

Tables 2 and 3 show the calculations of the relative performance factor

$\Delta_\omega = \frac{\omega}{\omega_{SISO}}$ performed according to formulas (6), (10)–(12) for different volumes of K of the input/output buffer.

ρ^*	$K = 1$	$K =$	$K = 3$	$K = 4$	$K=10$
1.25	1	1.33	1.58	1.69	1.76
0.85	1	1.33	1.59	1.69	1.78
4.5	1	1.17	1.22	1.22	1.22
2	1	1.23	1.45	1.45	1.5

Table 2. Relative performance for AISO

ρ^*	$K = 1$	$K =$	$K = 3$	$K = 4$	$K=10$
0.8	1	1.33	1.58	1.69	1.76
0.22	1	1.17	1.22	1.22	1.22
1.2	1	1.33	1.59	1.69	1.78
0.5	1	1.23	1.45	1.45	1.5

Table 3. Relative performance for SIAO

4. Conclusion

The closed network model has been developed on the basis of open-end pipeline queuing systems to track out the maximum operating pressure, detection the leak points, and monitor the parameters on a real-time basis. This enables faster analysis and data protection in the operation of modern industrial pipelines.

REFERENCES

1. Solution for the management, protection and control of pipeline leaks.
http://spb-xxi.ru/index.php?option=com_content&view=article&id=39&Itemid=41.
2. Vishnevsky V.M., Mukhtarov A.A., Pershin O.Yu. The problem of optimal placement of base stations of a broadband network to control a linear territory with a limitation on the amount of end-to-end delay. / Distributed computer and communication networks: control, computation, communications (DCCN-2020) [Electronic resource]: materials of XXIII International scientific conference, 14-18 Sept. 2020, Moscow / under total. ed. V.M. Vishnevsky, K.E. Samuilov; - M.: IPU RAN, 2020, pp. 148-155.
3. Modern Cryptography: Theory and Practice. .: Williams: 2005, p. 768.
4. Yermakov A.S. Models for evaluating performance in synchronous/asynchronous execution of input/output processes and information processing// KazNU Bulletin № 1 (60), 2009 . 16 – 23. Al-Farabi Kazakh National University.
5. Bisseling Rob H. Parallel Scientific Computation: A Structured Approach using BSP and MPI .Oxford, 2004 , p. 342 .
6. Hardy Y. and Steed Wily H. Classical and Quantum Computing: with C++ and Java Simulations (Paperback – Jan 18, 2002) –Birdhouse 2001, p.62
7. DiabWael William and Frazier Howard M. Ethernet in the First Mile: Access for Everyone. IEEE Standards Information Network (April 1, 2006), p. 450.

УДК: 681.324

Индустрия 4.0, система Mesh, модели и экосистема NG сетей

В. Л. Широков

Национальный исследовательский университет “МЭИ”, г. Москва

ShirokovVL@mpei.ru

Аннотация

Рассматривается концепция и новые технологии Четвёртой промышленной революции (Индустрии 4.0). Обсуждаются особенности Mesh систем доступа в сетях новых поколений NG (4G, 5G). Формулируются задачи, решаемые на разных этапах жизненного цикла сетей NG. Предлагаются решения, в которых сети NG с функциями Mesh-систем, будут экосистемой Индустрии 4.0. Обосновывается выбор базовой модели сети NG и Mesh-систем. Предлагается методика расчёта, оценки и выбора параметров Mesh-системы доступа для сети NG. В качестве основной метрики качества Mesh-систем определена временная задержка при передаче данных. Выбрана модель M/M/m кластера базовых станций для сетей NG. На основе модели разработана методика расчета задержек и эффективности Mesh систем в сетях NG как отношения стоимости оборудования к временным задержкам сети. В качестве мета модели сетей NG может использоваться теория Полуколец.

Ключевые слова: Индустрия 4.0, NG, M/M/m, Mesh, слой управления, слой данных, задержка

1. Введение

Во время Третьей промышленной революции для автоматизации различных производств использовалась фиксированная электроника и иногда информационные технологии.

Концепция Четвертой промышленной революции (Индустрии 4.0) заключается в следующем:

- массовое внедрение информационных технологий в промышленность;
- масштабная автоматизация бизнес-процессов и автоматики в производстве;
- использование преимуществ искусственного интеллекта и роботизации.

Безусловно, что Четвёртая промышленная революция должна решать непростые задачи. Однако и преимущества Индустрии 4.0 очевидны:

- существенное увеличение производительности труда;
- большая безопасность труда за счёт сокращения опасных рабочих мест;
- повышение конкурентоспособности.

И это, конечно, не полный список. А для того, чтобы использовать сети радиодоступа поколения NG с возможностями Mesh в качестве основы экосистемы Индустрии 4.0, необходимо внедрять эти радиосети в первую очередь в производство.

В течение всего жизненного цикла существования радиосистем, в процессе создания и внедрения радиосети формулируются и решаются следующие задачи:

- моделирование радиосети;
- выбор радиооборудования;
- проектирование радиосети;
- мониторинг и управление радиосетью.

Чтобы приступить к решению перечисленных задач, необходимо представить общую концепцию и инновационные технологии Индустрии 4.0, которые предстоит передать экосистеме NG.

2. Концепция и инновационные технологии Индустрии 4.0

Концепция Индустрии 4.0 состоит в цифровизации общества. Причём акцент должен быть направлен на автоматизацию промышленного производства. При этом цифровизация понимается в широком смысле, как цифровая трансформация, то есть кардинальное изменение существующей бизнес-модели и составляющих любого бизнеса, включая:

- создание новых промышленных продуктов;
- командную работу с использованием информационных технологий;
- автоматизацию производства, логистики, организацию взаимоотношений и обратной связи с клиентами.

Причём Индустрия 4.0 это не единичное внедрение какой-либо цифровой системы или инструментов. Это комплексная, кросс-функциональная бизнес-инициатива. В перспективе она должна перейти в автоматический режим. И она должна быть поддержана сотрудниками, одобрена топ-менеджментом и принята акционерами.

Ниже кратко представим инновационные технологии Индустрии 4.0, которые сети и экосистема NG позволят реализовать.

2.1. Большие данные (Big Data). Современные производственные системы должны собирать огромный объём больших данных (Big Data). Этих данных у компаний очень много: о сотрудниках, партнерах, продукции, клиентах. Их можно использовать с выгодой для бизнеса: улучшать продукцию, добиваться

роста продаж, снижать текучку кадров и т.д. Чтобы обрабатывать и анализировать большие данные, обнаруживать корреляции, используются специальные алгоритмы.

2.2. Финтех (FinTech). Это новый технологичный бизнес по предоставлению финансовых услуг: платежи, переводы, кредитование, краудфандинг.

2.3. Иншуртех (InsurTech) . Это новые технологии в области страхования, которые используют для оперативного общения с клиентами. И это могут быть специальные приложения, чат-боты, машинное обучение (ML), искусственный интеллект (AI). Например, машинное обучение учитывает спрос на продукцию, риски, обрабатывает претензии клиентов, обнаруживает случаи мошенничества.

2.4. Блокчейн (Blockchain) . Это технология распределенных баз данных, реестров, содержащих информацию, которая хранится в виде цепочек блоков обо всех транзакциях, проведенных участниками электронных сделок. При этом нет никакой централизованной организации, которая проверяла бы этот процесс.

В каждом из блоков конкретной цепочки записано определенное количество транзакций. Эта технология позволяет людям, не знающим друг друга, совместно использовать записи всех событий. Причём внутри этой системы невозможно скрыть или подделать какие-либо данные. Поэтому система является стойкой ко взлому, прозрачной в отношении коррупции, т.е. ещё и коррупционно стойкой.

2.5. Виртуальная реальность (Virtual Reality, VR) . Человеческий мозг запоминает всего 10% из прочитанного, 20% из услышанного, но зато 90% из того, что мы делаем. И это доказывает, что виртуальная реальность (VR) обеспечит изменение к лучшему наше обучение и различные тренинги, а в целом улучшит систему образования.

Под VR понимается искусственно созданная на компьютере виртуальная среда, доступ в которую осуществляется через так называемые иммерсивные («погружающие») устройства, создающие эффект нахождения человека в новой реальности. Этими устройствами могут быть шлемы, наушники, специальные перчатки. Такая виртуальная среда имитирует реальность. Пользователь взаимодействует с этой средой, погружаясь в неё полностью, например, при обучении, тренинге и т.д., а именно в созданный компьютером виртуальный мир, как в реальный, может «работать», манипулируя виртуальными объектами, выполнять действия, исследовать объект, обучаться «работе» в этой среде как в реальной.

2.6. Дополненная реальность (Augmented Reality, AR) . Данная реальность добавляет реальному миру дополнительные виртуальные слои. При этом человек по-прежнему как бы взаимодействует с физической средой. Но получает дополнительную информацию (графику, текст, анимацию, звук, видео, 3D-модели и т.д.), как будто от реальных устройств, а также – и от приложений

дополненной реальности. Цель этой реальности – предоставить пользователю более богатую аудиовизуальную обстановку, близкую реальной среде.

2.7. Искусственный интеллект (Artificial Intelligence, AI) . Искусственным интеллектом принято называть всё, что способно решать нерутинные задачи на уровне, близком к возможностям человеческого мозга. AI делает это почти мгновенно, а зачастую лучше и эффективнее. Эта технология собирает, обрабатывает, анализирует данные и принимает решения на основе анализа. Например, видеокамеры на дорогах отслеживают превышение скорости автомобиля, распознают номерной знак, идентифицируя владельца. Другое применение – в системах безопасности, обеспечивает обнаружение преступника в толпе, например, на вокзалах, в аэропортах, метро, других местах массового скопления людей.

2.8. Интернет вещей (Internet of Things, IoT) . Интернет вещей IoT способствует использованию в сети огромного количества устройств, подключая их и получая от них данные, позволяет собирать, анализировать, обрабатывать и передавать данные другим системам через приложения или через технические устройства.

IoT-системы, датчики, исполнительные механизмы работают в режиме реального времени и обычно состоят из сети умных устройств и облачной платформы. Они могут подключаться в корпоративную сеть, например, через локальную сеть типа Wi-Fi, персональную сеть типа Bluetooth или другие виды беспроводной связи. При этом точки доступа (AP) Wi-Fi и Bluetooth легко интегрируются в сети 4G и экосистему 5G.

2.9. Нейронная сеть (Artificial Neural Network) . Нейронная (искусственная) сеть – это последовательность ряда слоёв, каждый из которых состоит из искусственных «нейронов», выполняющих (каждый) свою функцию. Работа такой сети напоминает принцип действия нейронов человеческого мозга – отсюда её название. В процессе работы программа нейронной сети итерационно совершает огромное количество действий. В результате она определяет, какие значения искомых параметров будут приводить к наилучшему результату и решает задачу.

Тем самым нейронная сеть «самообучается», пытаясь выполнить задание. И сначала действует случайным образом. Но в процессе решения она получает обратную связь, итерационно приближаясь к решению и получая правильный ответ. При этом связи между «нейронами» сети, которые приводят к неправильному ответу ослабевают, а приводящие к правильному решению наоборот усиливаются. После множества проб, обратных связей, итераций, сеть обучается правильно решать задуманную задачу, формируя нужные нейронные пути.

2.10. Телемедицина (Telemedicine) . Телемедицина – это предоставление дистанционных медицинских услуг (например, мониторинг состояния пациента,

консультации, консилиумы). Она обеспечивает взаимодействие медицинских работников между собой и с пациентом при помощи телекоммуникационных систем и технологий.

2.11. Умная одежда (Wearable technology). Умная одежда может интерактивно взаимодействовать со средой, то есть считывать информацию об окружающей человека обстановке. Помимо специальных военных или медицинских костюмов, к умной одежде относят разные носимые гаджеты вроде умных часов, браслетов и фитнес-трекеров. Отдельная категория умной одежды – медицинские устройства, которые имплантируются в организм или носятся на теле, например, инсулиновые помпы, протезы конечностей, органов.

Таким образом, подходы Индустрии 4.0 основаны на внедрении информационных технологий в промышленность, автоматизации бизнес-процессов, внедрении АИ и других, перечисленных выше, технологий.

Главной целью Индустрии 4.0 является автоматизация производств, контроль, автоматизированное и, в перспективе, автоматическое управление технологическими и производственными процессами.

Ниже рассматриваются задачи, которые решаются с целью обеспечения цифровой трансформации, поддержания концепции и инноваций Индустрии 4.0, перечисленных выше.

3. Показатели качества сетей NG

Мобильная связь новых поколений NG (4G и 5G), по сравнению с предшествующими сетями 2G и 3G, обеспечивает существенно более высокие показатели производительности сети и её эффективности. Неплохие показатели были получены уже в сетях 4G за счёт совершенствования механизмов управления и маршрутизации трафика, приближающегося к сетям 5G.

Однако устройства и сети 4G по сравнению с 5G в целом всё-таки имеют недостаточные для производственных систем и Индустрии 4.0 масштабируемость, производительность, емкость и более высокое энергопотребление, чтобы массово использовать устройства IoT, количество которых растёт и будет расти экспоненциально.

Количество подключённых к сетям NG устройств IoT составит к 2025 году 55,7 млрд [1].

Ещё более высокими показателями качества обслуживания (QoS) трафика и пользователей (QoE) по сравнению с сетями 4G обладает экосистема 5G, а именно:

- пропускная способность на единицу площади в 1000 раз выше, чем у 4G;
- пиковая скорость передачи данных: 20 Гбит/с от базовой станции (БС) и 10 Гбит/с к БС, что в 100 раз выше, чем в 4G;

- минимальная скорость даже на границе соты 100 Мбит/с, а это пиковая скорость сети 4G;
- максимальное плотность подключений до 1 млн/кв.км, что в 100 раз больше, чем в 4G;
- максимально возможная полоса пропускания одного радиоканала до 1-2 ГГц;
- максимальная скорость передвижения абонентов, устройств – до 500 км/ч;
- временная задержка < 1 мс, по сравнению с 4G, которая > 10 мс;
- спектральная эффективность 30 бит/с/Гц от БС и 15 – к БС;
- доступность и надежность – 99,999%, покрытие – 100%;
- снижение энергопотребления – 90

В целях обеспечения перечисленных выше характеристик, в том числе в интересах Индустрии 4.0, для моделирования сетей NG необходимо:

- иметь адекватные математические модели;
- разработать соответствующие этим моделям методики;
- решать на каждом этапе жизненного цикла сети NG необходимый круг задач.

Далее рассмотрим выбранные модели, разработанные методики и решаемые задачи.

4. Сети NG, решаемые задачи, модели, методики

Очевидно, что для обеспечения Индустрии 4.0 должна использоваться высокоскоростная связь. Она нужна не только в интересах государственных органов, муниципальных предприятий и финансовых организаций. Однако в первую очередь такой связью должны охватываться промышленные предприятия и различные производства с целью их автоматизации для ускоренного развития производительных сил, включая производство продукции сельского хозяйства.

Наиболее приспособленными и удобными с точки зрения быстрого развертывания и масштабного охвата различных отраслей экономики и производств автоматизированными системами управления технологическими процессами, обеспечения обмена данными в реальном времени целесообразно использовать технологии беспроводной связи, которые уже пронизали другие сферы нашей жизни.

Поэтому для массового применения в Индустрии 4.0 предлагаются радиосети различного уровня: от глобальных сотовых сетей новых поколений NG (4G, 5G). С ними могут использоваться также локальные сети Wi-Fi, корпоративные WLAN и городские (муниципальные), или региональные, WMAN типа WiMAX

также относящиеся к новому поколению (4G). Кроме этого, могут использоваться персональные сети Bluetooth. Во всех перечисленных сетях может использоваться Mesh-система.

На макроуровне необходим учёт влияния на производственные системы глобальных сетей 4G и экосистем 5G, поскольку все эти сети могут использоваться в них. При этом наблюдается переход от устаревших сетей 2G и 3G к 4G и 5G. Например, число подключений IoT-оборудования через LTE-сети в 2021 году увеличилось на 24% благодаря более широкому использованию LTE-чипсетов.

Таким образом, объект исследования – беспроводные сотовые сети NG с использованием Mesh.

Класс исследуемых сетей: это многоканальные, мультисервисные, многофункциональные Mesh-системы обмена данными и радиодоступа к источникам информации.

Цели исследований: расчёт, анализ, выбор, а также мониторинг и автоподстройка параметров систем радиодоступа в сетях NG, оценка их эффективности.

Примечание. Под эффективностью понимается отношение выбранного параметра качества к стоимости системы или сети.

Параметром качества системы или сети может быть:

- скорость передачи данных и загрузка каналов связи системы;
- временные задержки при передаче данных.

При одинаковой стоимости качество выше, чем больше скорость передачи данных и лучшая загрузка. Однако при той же стоимости качество системы также выше, если меньше задержки при передаче, и в этом случае эффективность системы тоже выше.

Исходными данными при анализе таких систем и сетей будем считать следующие ограничения:

- марковские потоки поступления заявок на обслуживание;
- марковский закон обслуживания заявок;
- многоканальные обслуживающие узлы;
- общая среда передачи (радиоканал);
- ограниченные входные очереди.

Поэтому в качестве аналитической модели беспроводной сети выбрана система массового обслуживания $M/M/m$. Обоснованием выбора данной модели является минимальное количество соседних базовых станций (узлов сети NG) в кластере, как правило, 7. При этом количество активных узлов (базовых станций) в кластере можно считать марковским законом обслуживания заявок.

В качестве метрик при анализе параметров коммуникационных систем (КС) на данной модели, с учётом указанных выше ограничений, могут определяться следующие характеристики:

- стоимость оборудования;
- скорость передачи данных;
- скорость обслуживания заявок;
- задержки при передаче данных;
- длина очереди заявок на обслуживание;
- количество пользователей трафика в фоновом режиме;
- количество пользователей трафика в реальном времени.

Затем необходимо с целью оценки эффективности проектируемой сети, на выбранной выше модели, по разработанной и представленной ниже методике, последовательно и итерационно решать следующие задачи:

- 1) выполнить инженерное обследование и собрать исходные данные на проектируемую сеть;
- 2) выбрать оборудование и выполнить планирование радиосети (покрытие территории);
- 3) математически, имитационно и/или полунаатурным способом выполнить моделирование систем радиосети, то есть кластеров базовых станций;
- 4) выполнить расчёт, анализ, оценку, выбрать и/или уточнить параметры радиосети;
- 5) в случае необходимости уточнить исходные данные, сделать по ним корректировку модели, выполнить повторную итерацию, то есть прогнать модель по пп. 2-4 данной методики;
- 6) разработать проектную документацию и инсталлировать радиосеть.

Итак, во-первых, на модели (моделях) систем, во-вторых, на стадии опытной эксплуатации сети, в-третьих, к рабочему режиму функционирования сети необходимо подготовить (разработать и отладить) и использовать специальную систему мониторинга радиосети, контроля, управления её параметрами и конфигурирования оборудования. Однако эта задача очень комплексная, достаточно сложная, в данном докладе не анализируется и требует отдельного рассмотрения.

Отметим также особенности коммуникационных систем (КС), ориентированных на мобильных абонентов, роботов, машины, IoT. Эти особенности, а точнее ограничения и требования к системе, необходимо учитывать при проектировании и создании радиосетей. И они заключаются в следующем:

- необходимости передачи мультимедийной информации по радиосети в реальном времени;
- высокой плотности размещения активных узлов, точек доступа, базовых станций относительно друг друга;
- высокой пропускной способности, низкой задержке и малом времени отклика системы;

- определении, учёте и использовании местонахождения пользователей и обслуживаемых устройств.

Кроме этого, необходимо учитывать дополнительные требования с точки зрения пользователей и заказчиков к проектируемым КС, а именно:

- возможность обслуживания предприятий, общественных центров, кампусов, стадионов, как отдельных объектов;
- возможность оптимизации передачи видео- и аудио трафика, который должен передаваться в реальном времени;
- наличие и обслуживание шлюзов для специальных платформ интернета вещей (IoT);
- наличие пограничных служб мониторинга и визуализации состояния активных узлов сети (Edge Monitoring Services, EMS), а при необходимости, также служб пограничного контроля безопасности сети (Edge Security Services, ESS).

5. Проектирование, развертывание, управление сетей NG

Профессиональный интерес представляет методология системного проектирования, исследования, моделирования, развертывания, управления и модернизации КС новых поколений NG (4G и 5G).

Причём это нужно всем, начиная с проектировщиков до пользователей услуг этих сетей, и на всех стадиях жизненного цикла сетей NG:

- проектировщикам сетей 4G и 5G;
- исследователям, разработчикам и производителям телекоммуникационного оборудования;
- сервис-провайдерам Интернет и передачи данных, поставщикам контента, операторам связи;
- пользователям услуг Индустрии 4.0, включая обеспечение межмашинной связи M2M, интернета вещей IoT, промышленного PoT, других сервисов, а также – стартапам.

Отметим, что КС поколения 5G – это также высокоскоростные поезда, интеллектуальные системы здравоохранения, подключённые автомобили, умная городская инфраструктура, 90-% экономия электроэнергии, "нулевые" задержки, очень высокие скорости передачи данных.

В текущем десятилетии ожидается 100 кратное увеличение количества подключённых устройств. И в первую очередь это происходит за счёт всё большего распространения устройств IoT и PoT, а также устройств межмашинной связи M2M.

Работают устройства IoT и M2M в реальном времени. При этом IoT-система состоит, из следующих подсистем:

- специальной сети IoT умных устройств, подключённых через свою сеть к транспортной сети передачи данных более высокого уровня;
- специальной облачной программной платформы, работающей в среде той же транспортной сети.

На периферийном уровне сети для подключения IoT устройств можно использовать узкополосную (Narrow Band) сеть NB-IoT, например, LoRa (Long Range), обеспечивающую дальность действия от 8 до 30 км. Также может использоваться локальный Wi-Fi и/или даже персональный Bluetooth. Сети NG (4G, 5G) могут играть роль транспортной инфраструктуры.

В настоящее время устройства IoT уже могут подключаться непосредственно к сети 4G непосредственно в связи с производством LTE чипсетов.

Благодаря всё более массовому использованию интернета вещей IoT/ПоТ, уже в текущем десятилетии, как прогнозируется, и это показывает статистика, можно ожидать 1000-5000-кратный рост трафика. В 2021 году рост подключённых устройств IoT/ПоТ составил более 60%.

Чтобы справиться с этим взрывным увеличением трафика, ключевыми требованиями к сетям NG являются:

- общая пропускная способность сети;
- допустимая нагрузка на сеть;
- качество обслуживания QoS;
- качество восприятия QoE.

Пропускная способность сети определяется, как суммарный трафик, который сеть может пропустить и обработать.

Нагрузка на сеть – это реальный трафик, поступающий в сеть и обрабатываемый сетью.

Качество обслуживания (QoS) – это сетевые ориентированные параметры, такие как реальная интенсивность обслуживания и фактическая задержка.

Качество восприятия (QoE) – это сервис-ориентированный показатель качества обслуживания, определяемый как некоторый усреднённый балл, например, MOS (Mean Opinion Score) от 1 до 5 в телефонии, оценивающий разборчивость речи. В данном случае – субъективная оценка скорости обновления данных и задержки их получения.

Поэтому сетевая архитектура систем NG должна позволять сети радиодоступа (RAN) и ядру сети, например, 5GC (5G Core) модернизироваться и масштабироваться независимо друг от друга.

Ключевыми процессами мониторинга и управления, и задачами, решаемыми в рабочем режиме функционирования сетей NG, являются следующие:

- мониторинг, наблюдение, сбор и учёт статистических данных с целью улучшения управления сетью;
- визуализация и автоматизация процессов управления;
- масштабная автоматизация бизнес-процессов;
- использование искусственного интеллекта (AI) и методов машинного обучения (ML);
- использование технологий Data Science;
- обработка больших данных Big Data.

Все указанные задачи должны быть решены, и система должна обеспечивать автоматическую настройку всех процессов и параметров сети и автоматическое управление этими параметрами сети и их автоподстройку по системе SON самонастройка, самоорганизация, самооптимизация сети.

И мы с нетерпением ждём, когда перестанем говорить о пустом холодильнике, который сможет сам заказать сыр, когда он в нём закончится. Очевидно, что IoT в целом коренным образом изменит работу многих производств, автоматизирует их. А в идеале доведёт их до автоматического функционирования. И это кардинально изменит многие сферы человеческой деятельности.

6. Заключение

Рассмотрена концепция Индустрии 4.0 – цифровая трансформация, и не только в финансовой и житейской сферах, но и в промышленности.

Необходима модернизации существующих сетей 4G, приближение их к возможностям 5G, использование новых сетевых технологий: программно-определенной дезагрегации системы на программный и аппаратный слой, управления и данных, SDN принципов виртуализации NFV, Mesh и др. Отмечен переход от сеть-ориентированной модели качества обслуживания QoS к сервис-ориентированной модели качества восприятия QoE услуг.

Обоснован выбор модели M/M/m как базовой системы, представленной в виде кластера NG сети, и методики расчёта основного параметра качества в виде временных задержек в Mesh системе «кластер базовых станций – цепочка абонентских станций – пользователи», наиболее адекватной исследуемым сетям сотовой связи современных поколений NG (4G и 5G).

Литература

1. Как интернет вещей меняет бизнес [Электронный ресурс]. URL: <https://plus.rbc.ru/preview/61f031367a8aa959d425a2e1> (дата обращения: 01. 08. 2022).

2. Bhakta I. et al. Designing an efficient delay sensitive routing metric for IEEE 802.16 Mesh networks //arXiv preprint arXiv:1306.3903. – 2013.
3. Широков В. Л. Методология исследования и управления трафиком мульти-сервисных радиосистем обмена информацией //Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2016). – 2016. – С. 156-161.
4. Широков В. Л. Методология гибридного моделирования для оценки и выбора параметров радиосетей доступа //Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2017). – 2017. – С. 246-250.
5. Широков В. Л. МОДЕЛИ И МЕТОДЫ ГИБРИДНОЙ МЕТОДОЛОГИИ ДЛЯ ОЦЕНКИ И ВЫБОРА ПАРАМЕТРОВ КОНВЕРГЕНТНЫХ СЕТЕЙ РАДИОДОСТУПА 5G //Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2019). – 2019. – С. 398-404.
6. Shirokov V. L. Prerequisites and Methodology for Digital Transformation of 4G Networks into 5G Ecosystem //International Conference on Distributed Computer and Communication Networks. – Springer, Cham, 2020. – С. 158-168.

UDC: 004.738

FIREWALL SIMULATOR DEVELOPMENT FOR PERFORMANCE EVALUATION OF RANGING A FILTRATION RULES SET

A.Yu. Botvinko ¹ and K.E. Samouylov ¹

¹Peoples' Friendship University of Russia, Moscow, Miklukho-Maklaya, str.6, Moscow,
Russia

botviay@sci.pfu.edu.ru, ksam@sci.pfu.edu.ru

Abstract

This paper is written as a continuation of works devoted to solving the task of increasing the firewall performance in conditions of high heterogeneity and variability of the parameters of the filtered network traffic. The paper shows a simulation model that is intended for the evaluation of the major performance indicators of a firewall when ranging a filtration rule set. We've evaluated the effectiveness of the method for ranging a filtration rule set (it was developed earlier by the authors) for various parameters of the simulation model and different scenarios of network traffic behavior.

Keywords: firewall, ranging the filtration rules, network traffic, simulation model, queuing system, phase service, local approximation method.

1. Introduction

A firewall is a local or functional distributing tool that provides control over the incoming and/or outgoing information in the automated system (AS), and ensures the protection of the AS by filtering the information, i.e., providing analysis of the information by the criteria set and making a decision on its distribution.

Firewall is one of the major components of the network architecture. It ensures network security, including for special-purpose AS, the uninterrupted operation of which is critical for ensuring the security and defensive capability of any state. By using firewalls, you can solve such problems as preventing unauthorized access, and deleting, modifying, blocking, copying, providing and distributing information under protection. It's very important to ensure high performance of the firewall since there is a huge need to provide the stable operation of a critically important information infrastructures in the conditions of an avalanche-like growth in the volume of public network information flows, high heterogeneity and variability of network traffic

parameters, widespread use of multimedia protocols that are sensitive to the data transmission delays, and significant increase in the number of computer attacks.

Information filtering is executed by a certain rule set, determined in accordance with the security policy of the AS under protection. The filtration rules are a list of conditions under which further transmission of information is allowed or prohibited and a number of actions is performed by the firewall for registration and/or implementation of additional protective functions.

One of the major factors affecting the search time for rules, and therefore the firewall performance, is the order in which the filtration rules are arranged in sets that are linear lists of large dimensions. This is due to the fact that the search time for any rule corresponding to the filtered data, is in proportion to the number of checked rules. And the information flow filtering time that meets the conditions contained at the end of a large dimension set, will be much longer than the time required to filter data that meet the conditions contained at the beginning of a rule set. Therefore, the use of optimization methods for a filtration rule set is in demand for AS with a complex network architecture and large volumes of heterogeneous network traffic [1, 2].

The papers [1, 3] published earlier by the authors, describe the developed method for optimizing a filtration rule set (method for ranging the rules). This method takes changes in the parameters of information flows into account. An increase in the efficiency of traffic filtration can be provided by periodical ranging the filtration rules in descending order of their weights, obtained in accordance with the estimates of the parameters of the filtered information flows. A particularity of the developed approach is the use of the non-parametric method of local approximation (MLA) [4] to evaluate the parameters of filtered information flows. In the ranging process for a rule set, the current characteristics and dynamics of changes in the parameters of information flows are considered. At the same time, there is no need to select a parametric model that is acceptable for all evaluated parameters of information flows.

The implementation of MLA has provided the adaptability of the method, as well as a high response speed for changes in the parameters of filtered information flows thanks to the such specifics of MLA estimates as the use of:

- a local parametric model with a sliding region of parameter constancy and a controlled locality parameter that determines the dimension of the locality region;
- a special locality function to set the estimates of previous values when calculating the estimates of the parameters of filtered information flows.

In practice, the increased computational complexity of the developed approach when using modern microprocessors, doesn't have a significant affect on ensuring the stable firewall operation.

Therefore, a relevant task is to create a simulation model intended for the evaluation of the firewall performance with ranging the filtration rules. In contrast to earlier works [1, 2, 3] , the developed model allows us to obtain estimates for the firewall performance for different MLA parameters, as well types and scenarios of network traffic behavior.

2. Previous works. Firewall in 5G networks

In first papers devoted to solving the problem of improving the firewall performances, static methods to optimize the filtration rule set were used. These methods are deterministic and based on hardware solutions, heuristic algorithms and specialized data storage structures that don't depend on the parameters of filtered information flows [5]. Static optimization methods also include methods for finding and correcting inconsistency and redundancy of rules, which are defined as filtration rule configuration errors by a number of authors [6]. Other well-known optimization methods include early packet rejection optimization methods that create additional small dimension filters designed to early reject unwanted traffic before packets are verified by the main rule set. The use of such filters increases the firewall performance, since the rejection of unwanted traffic by the main rule set is executed after verifying all the rules of the set and takes the maximum possible filtering time [6, 7].

Currently, a relevant task is to solve the problem of increasing the firewall performance in fifth-generation mobile communication networks (5G/IMT-2020) [8, 9, 10]. A particularity of the use of firewall in 5G networks is high requirements for the duration of the data transmission delay to ensure ultra-reliable communication with low latency ULLRC (Ultra Low Latency Reliable Communication), as well as a very large number of filtration rules that are necessary to provide the secure functioning of massive computer-to-computer communications — Massive IoT/IIoT [11].

In paper [9], a new firewall is being developed for filtering traffic in 5G networks. It's intended for the analysis and filtration of network traffic in a specific segment (edge-to-core network segment of a 5G infrastructure). Field Programmable Gate Arrays (FPGA) and a P4 language (programming language) for programming packet routing rules are used to create the firewall.

The paper [11] is devoted to the development of a software firewall for 5G networks based on a software multi-layer switch with an open source text (OVS) designed to work in hypervisors and on computers with virtual devices. Developed firewall is intended to provide network traffic filtration for 5G Internet of Things devices. The

firewall performance makes it possible to set and use up to 1 million filtration rules for NB-IoT devices with 4 Gbps traffic.

The paper[10] is devoted to the efficient distribution of filtration rules between the firewalls within Internet of things (IoT). Automatic distribution, as well coordination of virtual firewalls is proposed. This is executed due to network function virtualization (NFV) management and orchestration (MANO) technology to protect NB-IoT mMTC communications. The major idea was to use NFV for efficient distribution of rules between firewalls based on VNFs to achieve scalability in terms of the number of IoT devices under control.

Also, it should be noted that most of the works devoted to the study of methods for optimizing a filtration rule set, don't imply the use of mathematical models to obtain estimates for the performance/efficiency indicators of the firewalls. The development of the firewall model with the possibility of changing the rule set will allow us to evaluate the effectiveness of optimization methods regardless of the specifics of the set rules and the characteristics of the hardware and software platform, as well as to analyze the dependence of performance indicators on the parameters of the rule optimization method.

3. Simulator development

The use of simulation modeling methods made it possible to eliminate restrictions on the type of the distribution function (DF) for the incoming packet flow and packet service for the earlier developed [1, 2, 3] firewall analytical models presented in the form of single-line queuing systems (QS), with a storage drive of limited capacity, heterogeneity of Poisson incoming flow, packet service with a phase-type service duration distribution function that depends on the order of the filtration rules. According to the Basharin-Kendall classification, QS data belong to $M_N/PH/1$ class. Hence, the implementation of the simulation model makes it possible to obtain estimates for the firewall performance when filtering various types and scenarios of network traffic behavior, as well as to evaluate the effectiveness of ranging the filtration rule set.

To create a simulation model that reproduces the filtration process for information flows and maintains the logical structure and sequence in terms of time, discrete-event modeling was chosen. Model traffic is given as a sum of Poisson flows for packets of various types and intensities. Packets incoming to the model are serviced in accordance with the FCFS principle. Only one packet can be served at a time, the other packets await for the service to start in a limited capacity storage drive. The packet service time distribution function (DF) is a phase-type function.

Since the purpose of modeling is to obtain estimates of the firewall performance indicators for various scenarios of the network traffic behavior and parameters of

the ranging method for the filtration rule set, we will describe a number of basic parameters that are necessary for modeling (see Table 3.1).

Parameter groups	Description of the parameters	Possible values
QS parameters	μ_0 — intensity of packet service during initial processing	double
	μ — intensity of service when checking whether a packet matches a specific (single) rule of the filtration rule set	double
	C — storage capacity of the system drive	uint8. Storage drive capacity is not less than 10% of the number of filtration rules in the set
Parameters of the filtration rule set	N — number of filtration rules in the set	uint8
	\mathbf{r}_1 — initial set of the filtration rules	$1 \times N$ uint8
Model traffic parameters	m — number of priority packets. Priority packets are packets with a high incoming intensity	uint8. The number of priority packets is less than the number of non-priority packets, $m < N - m$
	λ_1^{high} — initial intensity for the incoming priority packets	double
	λ_1^{low} — initial intensity for the incoming non-priority packets	double
	$f : (\lambda_{k,1}, \dots, \lambda_{k,1}) \rightarrow (\lambda_{k+1,1}, \dots, \lambda_{k+1,1})$	A function that sets the change in the intensity of the incoming packet flow after servicing the k -th packet group
Parameters of the ranging method	ω — calculation of filtration rule weights in accordance with the selected ranging method	uint8
Modelling time parameters	t — modelling time corresponding to the time of arrival of packets that make up one group	double
	t_0 — initial modelling time	double
	N_1 — number of packet groups in the data segment	uint8. $N_1 \geq 2$
	N_2 — number of data segments	uint8. $N_2 \geq 2$

Table 3.1. Simulation model parameters

The ‘QS parameters’ group corresponds to the parameters of the firewall mathematical model described in [1, 3] in the form of QS. The ‘Filtration rule set’ parameter group allows us to experiment with a randomly generated filtration rule set or, when it matters, with a predefined rule set that corresponds to a real one.

The ‘Parameters of the ranging method for the filtration rule set’ group is intended to set all the necessary parameters for the functioning of the ranging method for the filtration rule set. At the same time, for the convenience of obtaining comparative estimates of the firewall performance, some of the parameters are predetermined by the ‘calculation of filtration rule weights’.

‘Modeling time parameters’ set the time intervals, after which the system states change. Such values as the number of packet groups and data segments are necessary to implement the accumulation of experimental data and calculate the MLA estimates.

In most cases, writing an algorithm intended for modelling a complex process, in the form of a program, has significant difficulties and can lead to various errors. A large number of operators related to computational procedures and various support functions, make such a record obscure, and it becomes difficult to navigate in the modeling algorithm structure. That’s why the developed modeling algorithm for the firewall model with the ranging the filtration rule set reflects only the features of its structure, without unnecessary secondary details.

The algorithm is presented in the form of a pseudo-code similar, in terms of syntax, to the operators of the MATLAB system. Structurally, this modeling algorithm is divided into two algorithms — ‘Firewall simulation model’ and ‘Firewall functioning’.

Algorithm 3.1. ‘Firewall simulation model’ displays the filtration process for packet groups and data segments of the firewall, as well as the process of ranging the rule set and calculating the firewall performance indicators.

Algorithm 3.1. ‘Core algorithm’

Input parameters: $N, \mathbf{r}_1, m, \dots, N_1, N_2, t, f : (\lambda_{k,1}, \dots, \lambda_{k,1}) \rightarrow (\lambda_{k+1,1}, \dots, \lambda_{k+1,1})$

Output parameters: $V_k, W_k, U_k, Q_k, \rho_k$

- 1: Initialization: $k = 1, \Lambda = 0$
- 2: Calculating the probability of packet arrival: $\Lambda = m \cdot \lambda_1^{high} + (N - m) \lambda_1^{low}$
- 3: $\mathbf{P}(1 : m) = \lambda_1^{high} / \Lambda$.
- 4: $\mathbf{P}(m + 1 : N) = \lambda_1^{low} / \Lambda$.
- 5: for $j_2 = 1 : 1 : N_2$
- 6: for $j_1 = 1 : 1 : N_1$
- 7: k -th launch of the Simulink model for the t time duration
(Algorithm 3.2. ‘Firewall algorithm’)
- 8: $S_k \leftarrow$ get the values of the vector containing the numbers
of the rules that matched the served packet from
the Simulink model for the N_1 -th group of the N_2 data segment
- 9: Calculate x_i^k – the number of packets that match
the r_i^k rule for $i=1:1:N$
- 10: $x_i^k = \text{sum } (S_k(:) = i)$

```

11:      end
12:      Calculate  $\delta$ 
13:      if  $j_2 =$  the first data segment
14:           $\hat{x}_k^i = 0, i = 1, \dots, N, \mathbf{p}_k = [p_1^k, \dots, p_N^k] = \mathbf{0}$ 
15:      else
16:          calculate  $\hat{x}_{k+1}^i, i = 1, \dots, N$ 
17:          Set rule weights:  $\mathbf{p}_k = [p_1^k, \dots, p_N^k] = [\hat{x}_{k+1}^i, \dots, \hat{x}_{k+1}^N]$ .
18:      end
19:      Calculate:  $\gamma_j(k) \in \{1, \dots, N\}, j = 1, 2, \dots, N$ 
20:      Get a ranged set using  $\mathbf{r}_k$ :  $\mathbf{r}_{k+1} = [r_{\gamma_1(k)}^k, \dots, r_{\gamma_N(k)}^k]$ 
21:       $V_k, W_k, Q_k \leftarrow$  get values of the performance indicators
         from the Simulink model
22:      Calculate performance indicators:  $U_k, \rho_k$ 
23:      Set the intensity of incoming packets for the following packet group:
          $\lambda_{1,k+1}, \dots, \lambda_{N,k+1} = f(\lambda_{1,k}, \dots, \lambda_{N,k})$ 
24:       $k = k + 1$ 
25:      End
26:  End

```

Algorithm 3.2. ‘Firewall algorithm’ displays only the main procedures that ensure the functioning of the earlier considered QS with phase-type packet service. At the same time, the sequence of the processes of packet arrival, waiting, and servicing, and the transitions between blocks implementing the QS, are not shown in the algorithm.

Algorithm 3.2. ‘Firewall functioning’

Input parameters: $N, \mathbf{P} = [p_{1,i}, \dots, p_{1,N}], C, \mu_0, \mu, \Lambda$.

Output parameters: V_k, W, Q_k, S_k

- 1: Generate a number with a equal distribution on the interval $[0, 1]$ $a \sim U(0, 1)$
- 2: Calculate the arrival time for the packets; $dt = \Lambda^{-1} \ln(1 - a)$
- 3: Generate a number with an equal distribution on the interval $[0, 1]$ $b \sim U(0, 1)$
- 4: Get the type of the packet that incomes the QS:
 $e = \text{sum}(b \geq [0, \sum_{i=1}^1 p_{1,i}, \dots, \sum_{i=1}^N p_{1,i}])$
- 5: Generate a packet of $etype$ using the ‘Entity Generator’ block
 of the Simulink model
- 6: Generate a final packet queue with length using the ‘Entity Queue’ block
 of the Simulink model
- 7: With a free ‘Entity Server’ block

- determine the i -th number of the rule corresponding to the incoming packet of e type;
 - calculate the service time (v_e) using the packet service function $g : (e, \mu_0, \mu, \mathbf{r}_k, N) \rightarrow [v_e, i]$.
- 8: Over the v_e time, provide the packet service in the ‘Entity Server’ block of the Simulink environment
- 9: Complete servicing the packets in the ‘Entity Terminator’ block
- 10: Proceed to servicing and generating the next packets
- 11: After completion of t interval, using the statistics of the ‘Entity Queue’ and ‘Entity Server’ blocks and the Simulink model, transfer the following values to the Algorithm 3.1 — ‘Core algorithm’: V_k, W_k, Q_k, S_k .

For the software implementation of the modeling algorithm, we used the Simulink environment for dynamic interdisciplinary modeling of complex technical systems with the SimEvents library of discrete states that uses the apparatus of the theory of queues and queuing systems.

A particularity of the Simulink environment is a high degree of integration into the MATLAB matrix calculation system. This makes it possible to launch a model from a MATLAB script file, get the parameters from it, as well as send the modelling results back to the working environment. It allowed us to use built-in mathematical algorithms and tools when implementing the modeling algorithm and processing the experimental results. With the MATLAB system, the scalability of the model was also provided. It became possible to add and expand the functionality of the simulation model without significant changes in the existing project architecture.

4. Case study and performance analysis

The purpose of the experiment was to evaluate the firewall performance when ranging the filtration rules by the method proposed in papers [1, 3] for various traffic types and MLA parameters.

When modeling, two types of system operation conditions were under consideration. They correspond to the normal functioning of the system and the firewall functioning the under overload conditions.

The normal conditions mean such a ratio of intensities of network packet arrival, intensities of packet service and storage drive capacity that doesn’t lead to system operation with a load being $\rho > 0.9$.

The functioning of the system under overload conditions was considered for incoming traffic with a harmonically changing intensity of the incoming flows depending on the number of the packet group being serviced, for which the system operates with

overload or is close to the packet loss limit. There are studies [12, 13] of similar QS models with harmonic fluctuations of the incoming flow intensity, where the analysis of the stability of the packet queue characteristics was performed. Approximately, these conditions can be considered as the modeling process for the firewall operation with overloads, for example, when implementing DDoS attacks.

Depending on the values of the MLA parameters and the selected weights, the following methods for ranging the filtration rule set were considered:

- 1) Ranging with adaptive δ was performed in accordance with the 1st order MLA estimates with an adaptive locality parameter δ .
- 2) Ranging by least-square method (LSM) was executed in accordance with the 1st order MLA estimates at $\delta = \infty$. The estimates obtained are LSM estimates.
- 3) Ranging without levelling out was executed in accordance with the 1st order MLA estimates at $\delta \rightarrow 0$.
- 4) No ranging . The weights of the rules are $p_i^k = 0, i = 1, \dots, N$.

To evaluate the effectiveness of the ranging method (see section 2.4), the average values — $M(\Delta Z^\omega)$, $M(\Delta \delta_Z^\omega)$ — and maximum values — $\max(\Delta Z^\omega)$, $\max(\Delta \delta_Z^\omega)$ — of absolute and relative errors of performance/efficiency indicators calculated for packet groups without applying the $Z^0 = \{V_i^0, W_i^0, U_i^0, Q_i^0, \rho_i^0\}$ rule ranging, and $Z_i^\omega = \{V_i^\omega, W_i^\omega, U_i^\omega, Q_i^\omega, \rho_i^\omega\}$ performance/efficiency indicators calculated for packet groups after applying the ranging method are used with being the ranging method and n being the number of serviced packet groups since the first ranging.

In the experiment with the normal functioning of the system, the following initial data were chosen to calculate the performance indicators: the number of filtration rules in the set — $N = 1000$, the storage drive capacity in the system — , which is 10% of the number of rules, the time for the packet initial processing — $\mu_0^{-1} = 2.7 \cdot 10^{-3}$ [ms], the time for checking one rule — $\mu^{-1} = 5 \cdot 10^{-5}$ [ms]. The intensity of incoming packets (see Figure 4.1) changes after servicing each packet group so that $\Lambda_{k+1} = \Lambda_k + X_{k+1} \cdot \Lambda_k$ — with X_k being a random variable equally distributed on the interval of [-0.1, 0.2]. The initial the intensity value is $\Lambda_1 = 20$ [ms^{-1}], the number of packet groups is $k = 60$, the time interval of the system operation is $T = 60$ [s] The packet flow that incomes to the system is the sum of the Poisson packet flows. The number of flows with a high intensity is 15% ($m = 150$), their corresponding intensities are $\lambda_{i,k} = \lambda_k$, $i = 1, \dots, m$. The intensities of the remaining flows are $\lambda_{i,k} = \lambda_k$, $i = m + 1, \dots, N$. The maximum value of the total intensity is 31.21 [ms^{-1}], the average value is 19.53 [ms^{-1}]. The maximum intensity of the priority packet flow is 0.0975 [ms^{-1}], for non-priority packets it is 0.0195 [ms^{-1}].

In the experiment with a harmonically changing intensity, depending on the number of packet group being serviced, the initial data were equal to those used when

modeling the traffic filtration process with normal system functioning conditions, except setting the intensity of packet arrival — the intensity changes after servicing each packet group, so $\Lambda_k = \frac{a_0}{2} + \sum_{l=1}^2 \left\{ a_l \cos \left(\frac{2\pi}{T} lk \right) + b_l \sin \left(\frac{2\pi}{T} lk \right) \right\}$ with $T = 30$, $a_0 = 110$, $a_1 = 1$, $a_2 = 9$, $b_1 = -4$, $b_2 = -4$, $k = 1, \dots, 60$. The maximum value of the total intensity is $31.21 \text{ [ms}^{-1}]$, the average value is $19.53 \text{ [ms}^{-1}]$. The maximum intensity of the priority packet flow is $0.0975 \text{ [ms}^{-1}]$, the maximum intensity for the non-priority packet flow is $0.0195 \text{ [ms}^{-1}]$.

5. Analysis of firewall performance indicators

The estimates for the firewall performance indicators (with ranging the rules), obtained during modelling under system overload conditions, are given in Figures 5.1-5.4. The maximum values of absolute and relative errors of performance indicators for various MLA parameters are given in Table 5.1.

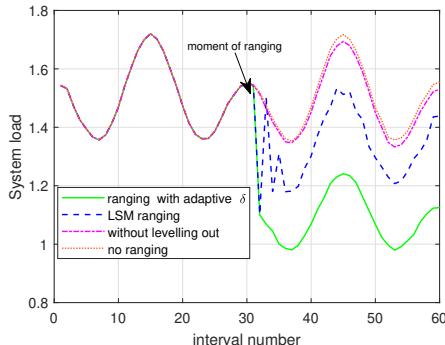


Fig. 5.1. System load

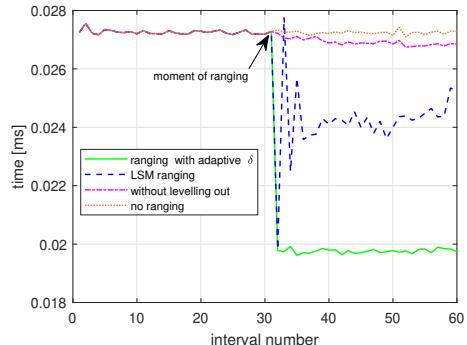


Fig. 5.2. Average service time

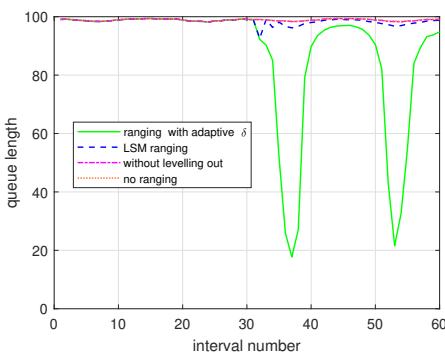


Fig. 5.3. Average queue length

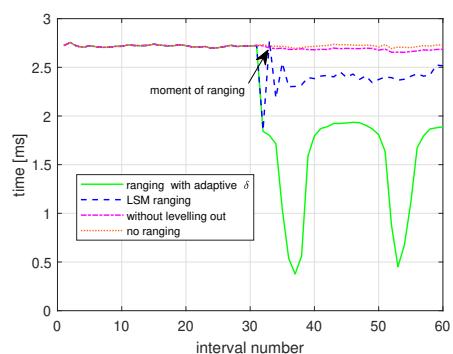


Fig. 5.4. Average residence time in the system

Optimization of the filtration rule set led to a decrease in the load within the system (from the moment the first ranging was launched) — on average from 1.539 to 1.103. Figure 5.1 shows the load graphs in the system for various MLA parameters. The maximum load in the system without ranging was 1.72. The change in the load (showed in Figure 5.1) corresponds to a change in the total intensity with a significant decrease in the load when ranging with adaptive δ and LSM ranging.

When the filtration rule set is optimized, there is a decrease in service time from the moment of the first ranging. Figure 5.2 shows the average service time for different MLA parameters. Qualitatively, the obtained results do not differ from those obtained for the system operation without overloads — a change in load doesn't change the service time, while a significant decrease in service time is obtained when ranging with adaptive δ and LSM ranging.

As shown in Figure 5.3, ranging the filtration rule set significantly decreases the residence time from the moment of the first ranging. Similar to the case without overloads, the system eventually turns into an established mode when servicing one packet group. It should be noted that despite the fact that the load in the system is more than 1, the average packet residence time in the network doesn't become infinitely large because of the limited storage drive capacity.

Ranging with adaptive δ and LSM ranging significantly reduce the queue length, as demonstrated in Figure 5.4. The best results were obtained for ranging with adaptive δ (see Table 5.1).

Modelling the process of traffic filtration with a harmonically changing intensity of the incoming flow showed the results as follows:

- 1) The use of ranging allowed us to reduce the average service time, which provided significant decreases in the system load and in the values of all performance indicators.
- 2) On average, ranging with adaptive δ demonstrated higher performance/efficiency compared to other methods.
- 3) For ranging with adaptive δ :
 - (a) Decrease in the system load is 0.39 (26.5%).
 - (b) Decrease in the average packet service time is 0.007 ms (26.5%).
 - (c) Decrease in the average packet residence time is 1.16 ms (42.6%).
 - (d) Decrease in the average waiting time for service is 1.15 ms (42.8%).
 - (e) Decrease in the average queue length is 22.23 (22.6%).

Compared to the results obtained for modelling the traffic filtration process under normal operating conditions, the absolute errors of performance indicators have higher values, except for the average service time, since it doesn't depend on the system load.

System load				
Ranging type, ω	$M(\Delta\rho^\omega)$	$\max(\Delta\rho^\omega)$	$M(\delta_\rho^\omega)$	$\max(\delta_\rho^\omega)$
With adaptive δ	0.397436	0.952226	26.54746	56.26077
Least-square method (LSM)	0.161254	0.834968	10.76459	54.9313
Without levelling out	0.017484	0.049596	1.161192	3.528732
Average service time				
With adaptive δ	0.007232	0.01541	26.54725	56.26145
Least-square method (LSM)	0.002932	0.015016	10.76413	54.93123
Without levelling out	0.000316	0.00096	1.161012	3.523324
Average residence time in the system				
With adaptive δ	1.157548	4.634046	42.65268	172.0129
Least-square method (LSM)	0.316624	1.755434	11.64846	64.26404
Without levelling out	0.032942	0.103576	1.211598	3.825138
Average queue length				
With adaptive δ	22.23406	161.2538	22.55491	163.9302
Least-square method (LSM)	1.052322	12.94445	1.065178	13.06728
Without levelling out	0.053542	0.32374	0.054194	0.32865

Table 5.1. Performance indicators

6. Conclusion

In this paper, a simulation model has been created to evaluate the main firewall performance indicators when ranging a filtration rule set. An estimate of the effectiveness of the ranging method for the filtration rule set is obtained for various parameters of the simulation model, as well as for various scenarios of network traffic behavior and MLA parameters. On average, for normal conditions and system overload conditions, ranging with adaptive δ demonstrated higher efficiency compared to other methods. This can be explained by a smaller approximation error of MLA with adaptive δ compared to the LSM evaluation [4]. Thus, the results of this work confirm the assumption about the increase in the firewall performance due to the use of the method of ranging a rule set that is adaptive to changing parameters of information flows, including the case of DDOS attacks

REFERENCES

1. Botvinko, A. Evaluation of firewall performance when ranging a filtration rule set / A. Botvinko, K. Samouylov // Discrete and Continuous Models and Applied Computational Science. – 2021

2. Botvinko, A. Evaluation of the firewall influence on the session initiation by the sip multimedia protocol / A. Botvinko, K. Samouylov // Discrete and Continuous Models and Applied Computational Science. – 2021. – № 3. – P. 221–229. – DOI 10.22363/2658-4670-2021-29-3-221-229.
3. Botvinko, A. Firewall Simulation Model with Filtering Rules Ranking / A. Botvinko, K. Samouylov. In: V. Vishnevskiy, K Samouylov., D. Kozyrev (eds) // Communications in Computer and Information Science : Distributed Computer and Communication Networks, DCCN 2020. – Cham : Springer, 2020. – Vol 1337. – P. 533 545.
4. Hardle, W. Applied nonparametric regression. // Cambridge university press. – 1990.
5. Zhu, Y. C. Optimization design and implementation of gateway based on firewall for access control / Y. C. Zhu // proceedings 6th International Conference on Information Science and Technology, ICIST 2016, 6 8 May 2016. – Dalian : IEEE, 2016. – P. 100 104. – DOI 10.1109/ICIST.2016.7483393.
6. Al Shaer, E. Automated firewall analytics: Design, configuration and optimization / E. Al Shaer. – Cham, Heidelberg, New York, Dordrecht, London. : Springer, 2014. – 132 p. – DOI 10.1007/978 3 319 10371 6.
7. Bagheri, S. Dynamic Firewall Decomposition and Composition in the Cloud / S. Bagheri, A. Shameli Sendi // IEEE Transactions on Information Forensics and Security, 2020. – Vol. 15. – P. 3526-3539. – DOI 10.1109/TIFS.2020.2990786.
8. Hybrid Tree rule Firewall for High Speed Data Transmission / T. Chomsiri, X. He, P. Nanda, Z. Tan // IEEE transactions on cloud computing, 2016 vol. – 8. – №. 4. – P. 1237-1249.
9. Hardware-accelerated firewall for 5g mobile networks / Ricart-Sanchez R. [et al] // IEEE 26th International Conference on Network Protocols (ICNP), 2018. – P. 446–447.
10. Netfpga-based firewall solution for 5g multi-tenant architectures / Ricart-Sanchez R. [et al] // 2019 IEEE International Conference on Edge Computing (EDGE), 2019. – P. 132–136.
11. Towards automatic deployment of virtual firewalls to support secure mMTC in 5G networks / Salva-Garcia P. [et al] // IEEE INFOCOM. 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2019. – P. 385–390.
12. Highly-scalable software firewall supporting one million rules for 5G NB-IoT networks / Escolar A. [et al] // ICC 2020-2020 IEEE International Conference on Communications (ICC), 2020. – P. 1–6.
13. Markova, E. Queuing system with unreliable servers and inhomogeneous intensities for analyzing the impact of non stationarity to performance measures

- of wireless network under licensed shared access / Markova, E. // Moscow : Mathematics. 2020. – Vol. 8. – №5. – DOI 10.3390/math8050800.
14. Modeling and analyzing licensed shared access operation for 5G network as an in-homogeneous queue with catastrophes / I. Gudkova, A. Korotysheva, A. Zeifman [et al] // proceedings 8th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2016, 18–20 Oct. 2016. – IEEE: Lisbon, 2016. – Vol. 2016. – P. 282–287. – DOI 10.3390/math8050800.

УДК: 001.57

Численное исследование вероятности идентификации RFID-метки с помощью RFID-считывателя, размещенного на БПЛА

В.Л. Абрамян¹ and А.А. Ларионов²

¹Московский физико-технический институт, Институтский пер. 9, Долгопрудный,
Россия

²Институт проблем управления РАН, ул. Профсоюзная 65, Москва, Россия

abramian@phystech.edu, larioandr@gmail.com

Аннотация

Современное развитие беспилотных летательных аппаратов (БПЛА) открывает большие возможности для их применения в различных областях промышленности и хозяйства. В данной статье исследуется совместное использование технологии радиочастотной идентификации (RFID) и БПЛА, а именно рассматривается считывание пассивных RFID-меток со считывателя, закреплённого на БПЛА. Для этого была разработана модель канала связи между RFID-устройствами, а также проведены численные эксперименты с помощью дискретно-событийного моделирования. Была вычислена оценка вероятности успешного обмена данными между считывателем и метками при разных входных данных: параметров протокола RFID, состояния окружающей среды, режима полёта БПЛА и т. п. Показано, что при определённых настройках RFID-считывателя возможно получение данных от метки с вероятностью свыше 0,95 при скорости полета до 30 км/ч

1. Введение

Технология RFID разрабатывалась как простой и эффективный способ бесконтактной идентификации объектов для широкого круга задач в области логистики, транспорта промышленного производства, контроле доступа на предприятиях и во многих других областях. Существует несколько реализаций технологии RFID, различающихся диапазонами радиочастот, протоколами связи и, как следствие, дальностью работы, скоростью и предельным объемом получаемых данных. В настоящей статье исследуется технология UHF RFID, работающая в диапазоне 860 - 960 МГц и описываемая стандартом EPC Class 1 Gen.2 [1].

Особенностью RFID является то, что она состоит из активного считывателя и пассивных меток, не имеющих собственного источника питания. Работа контроллера и отправка ответов от метки происходят за счёт использования

энергии электромагнитного поля, созданного считывателем. Метка использует метод обратного рассеяния [2], модулируя постоянную составляющую сигнала считывателя с помощью изменения коэффициента отражения. Благодаря этому метка может обходиться без встроенного источника энергии, но дальность её работы существенно ограничена. Для типичных меток, рассматриваемых в работе, дальность связи не превышает 12-15 метров.

В настоящее время ведётся разработка меток, совмещённых с датчиками, которые способны не просто хранить, но и передавать данные, полученные со встроенным сенсором. Это делает возможным использовать технологию RFID для сбора данных с сенсорных полей, идентификации автомобилей на парковке, поиска объектов на открытых складах. В данном случае инвентаризацию возможно проводить с помощью считывателя, закреплённого на БПЛА [3, 4].

Совместному использованию беспилотных летательных аппаратов и RFID посвящено большое количество исследований. Например, в работах [5, 6] рассматриваются различные модели, описывающие связь считывателя на БПЛА с метками. В [7] рассматривается применение считывателей, закреплённых на различных беспилотных транспортных средствах (квадрокоптер и машинка). В статье [8] рассматривается вопрос взаимодействия технологии БПЛА и RFID для задачи целостного автоматизирования сельскохозяйственной деятельности. В работах [9, 10] рассматривается вопрос определения местоположения пассивных меток с помощью считывателя, закреплённого на БПЛА.

Для эффективной работы RFID на беспилотном летательном аппарате требуется найти такие параметры полёта и протокола, благодаря которым станет возможно считывание с требуемой вероятностью. Для этой цели в работе строится модель канала для расчёта затуханий, а также имитационная модель, симулирующая полёт БПЛА вдоль стены и обмен сообщений по протоколу EPC Class 1 Gen.2, оценивающие искомую вероятность успешного обмена данными. Формально, в статье решается задача вычисления оценки вероятности $P(N_e, N_D; C, D, L)$, где N_e – число бит идентификатора EPCID, N_D – число бит данных, C – состояние канала, D – параметры полёта, L – параметры протокола RFID. При этом рассматривается частный случай равномерного и прямолинейного полёта БПЛА вдоль бетонной стены (см. рис. 1), что позволяет обобщить полученные результаты для случая движения считывателя не только по открытым участкам поля, но и по его границе, окружённой забором. Модель учитывает прямой и отражённый от стены луч, а также эффект Доплера, возникающий из-за движения считывателя. В общей сложности учитывается более пятидесяти параметров.

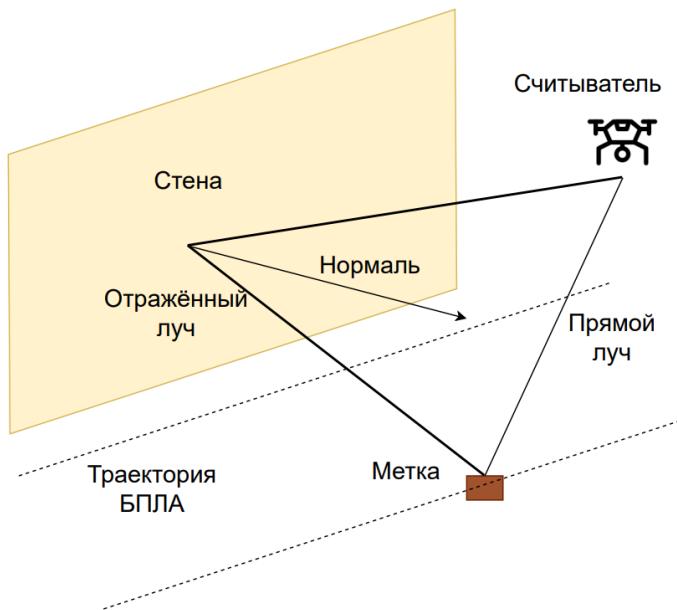


Рис. 1

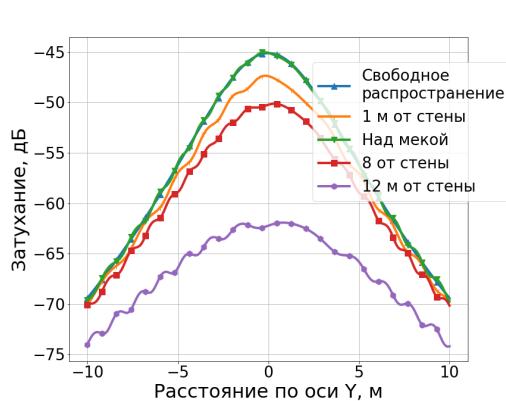
2. Модель канала

Затухание сигнала в результате двухлучевого распространения при отражении от бетонной стены было вычислено по формуле (1), где A_{pl} – искомые потери, $r(t)$ – используемая модель сигнала, λ – длина волны, d_i – длина пути i -го луча, N – количество лучей (здесь $N = 2$), R_i – коэффициент отражения от поверхности для i -го луча, Γ_i – множитель, учитывающий диаграммы направленности антенн считывателя и метки, v – скорость полёта БПЛА, ϕ_i – угол прихода волны относительно направления движения приемника, t – время, k – волновое число, равное отношению 2π к длине волны λ , h_i – затухания, вызванные различными факторами, $s(t)$ – однолучевая компонента или "копия" суммарного сигнала, τ_i – i -я задержка для i -й копии сигнала, ν_i – доплеровский сдвиг частоты i -й компоненты сигнала.

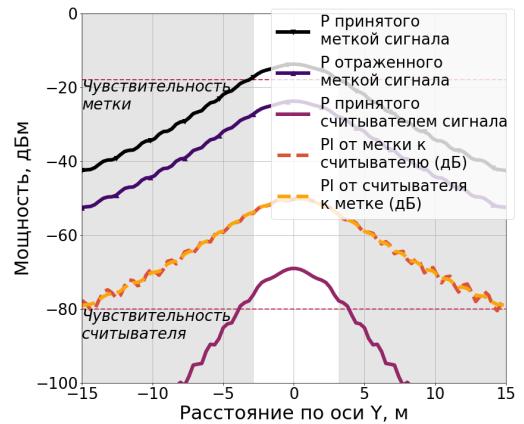
$$A_{pl} = |r(t)|^2 = \left(\frac{\lambda}{4\pi} \right)^2 \left| \sum_{i=0}^N \frac{R_i \Gamma_i}{d_i} e^{-jk(d_i - vt \cos \phi_i)} \right|^2, \quad (1)$$

$$r(t) = \sum_{i=0}^N h_i s(t - \tau_i) e^{j\nu_i t}, \quad (2)$$

На Рис. 2 а) представлены результаты моделирования потерь в канале для разных расстояний пролёта считывателя относительно стены. В данном случае метка оставалась на расстоянии 5 м от препятствия. Для наглядности изображена кривая, описывающая потери в случае свободного распространения сигнала. В таблице 1 приведены значения остальных параметров затухания, которые учитываются для изображения кривых на Рис. 2 б), на котором показан расчёт бюджета канала. В легенде рис. б) P - это мощность, а Pl - потери.



а) Потери в канале от считывателя к метке



б) Бюджет соединения

Рис. 2. Характеристика канала

BER был вычислен по формуле 3, где M - тип модуляции ответов метки, SNR - отношение уровня сигнала к шуму [11].

$$BER = \frac{1}{2} - \frac{1}{\sqrt{1 + \frac{2}{M*SNR}}} + \frac{2}{\pi} \frac{\arctan(\sqrt{1 + \frac{2}{M*SNR}})}{\sqrt{1 + \frac{2}{M*SNR}}} \quad (3)$$

3. Имитационная модель

Разработанная имитационная модель описывает работу протокола RFID, симулирует полёт БПЛА со считывателем и вычисляет вероятность успешного обмена данными. В ядре модели содержится приоритетная очередь*, содержащая все генерируемые события, например, включение считывателя, появления в его зоне видимости метки, обновление положения считывателя относительно метки,

*Приоритет по времени наступления события

Параметр	Значение
Мощность, излучаемая считывателем, $P_t^{(r)}$	31,5 дБм
Усиление антенны считывателя, $G^{(r)}$	8 дБи
Усиление антенны метки, $G^{(t)}$	2 дБи
Чувствительность метки, $P_s^{(t)}$	-18 дБм
Потери на поляризации, A_{pol}	-3 dB
Потери на модуляции на метке, A_m	-10 dB
Потери в кабеле, A_{cab}	-2 dB

Таблица 1. Параметры затухания, не зависящие от расстояния между считывателем и меткой

выход метки из зоны видимости, пересчёт параметров канала и т.д. То есть изменение состояния модели может произойти только в результате наступления того или иного события. Среди параметров протокола RFID следует выделить наиболее важные и которые варьировались в экспериментах: Tari – длительность символа data-0 (нуля) считывателя, и может принимать значения в 6,25, 12,5, 18,75, 25 мкс. Число M - тип кодирования ответов метки принимает значения: 1, 2, 4 и 8, которые соответствуют кодировке FM0 и кодам Миллера с 2, 4 или 8 символами на бит соответственно.

Имитационная модель была использована для проведения численных экспериментов с разными параметрами, после чего были сделаны выводы о режиме полёта, обеспечивающим необходимую вероятность успешного считывания меток.

На Рис. 3 и 4 представлены результаты двух численных экспериментов для расчёта зависимости вероятности успешного чтения. В первом эксперименте исследовалась зависимость вероятности от скорости БПЛА при различных параметрах кодирования команд считывателя (значение длительности символа Tari) и ответов меток (тип кодирования FM0 или коды Миллера с различным числом символов на бит). Во втором эксперименте исследовалась зависимость вероятности успешного чтения от высоты полёта для различных дистанций БПЛА от стены. Константы, оказывающие существенное влияние на эксперимент для рис. 3: чтение 64 слов, высота полёта 5 м, расстояние от стены 5 м, мощность считывателя 29 дБм; для рис. 4: скорость 30 км/ч, чтение 64 слова, расстояние от стены 2 м, мощность считывателя 29 дБм, кодирование ответов меток: Миллер-2, Tari = 12,5 мкс

Из приведённых численных экспериментов можно получить информацию о параметрах, удовлетворяющих требуемой величине вероятности успешной инвентаризации. Например, при кодировании ответов FM0 достаточно надежного

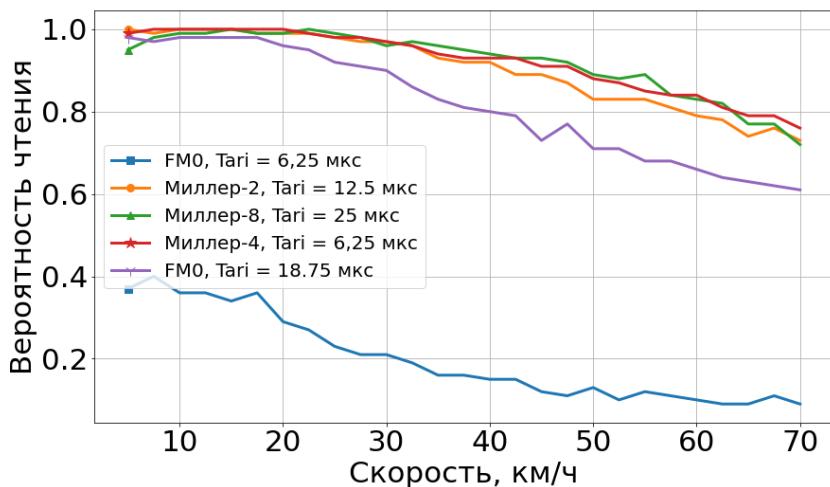


Рис. 3

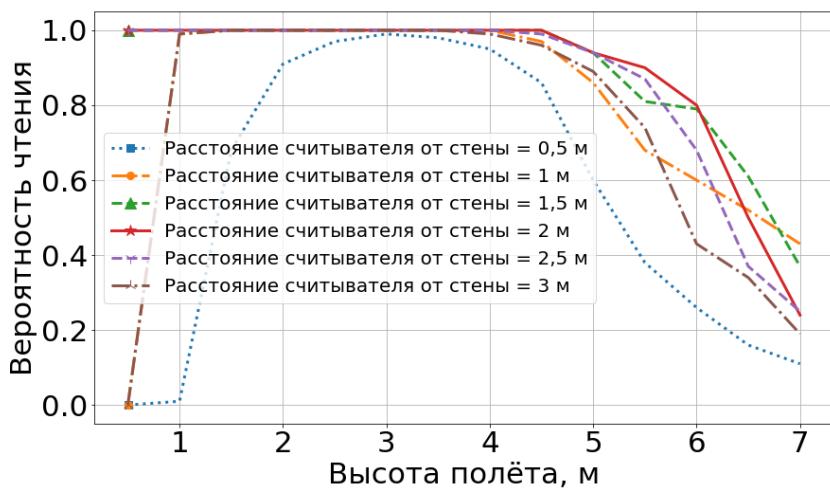


Рис. 4

чтения можно добиться только при больших значениях T_{ari} и невысокой скорости, до 20 км/ч; при скорости до 30 км/ч и высоте полета 5 метров большинство настроек системы, в которых используется более одного символа на бит, позво-

ляют получить вероятность чтения не ниже 0,95; расстояние от считывателя до стены оказывает существенное влияние. В частности, если считыватель оказывается значительно ближе к стене, чем метка, вероятность чтения данных очень сильно зависит от высоты. Например, при нахождении метки в 2 метрах от стены, пролёт считывателя на расстоянии 0,5 метра со скоростью 30 км/ч, и работе считывателя с мощностью 29 дБм, надежное чтение 64 слов данных возможно на высотах от 2 до 4 метров. Если же считыватель пролетает точно над меткой, то при тех же входных параметрах вероятность чтения 0,95 достигается на всех высотах до 4,5 метров.

4. Заключение

В статье была представлена модель канала между RFID-считывателем, расположенным на БПЛА, и RFID-меткой, размещенной на поле. Модель учитывает наличие боковых препятствий и эффект Доплера, возникающий из-за движения считывателя. Также была разработана имитационная модель для исследования вероятности чтения данных с метки, позволяющая учесть параметры полёта и протокола RFID.

Литература

1. EPC™ Radio-Frequency Identity Protocols Generation-2 UHF RFID Standard. Specification for RFID Air Interface Protocol for Communications at 860 MHz – 960 MHz. Release 2.1. — EPCCglobal, 2018. — C. 157
2. Finkenzeller K. RFID технологии. — 2003. — DOI: 10.1002/0470868023.
3. Yang J. H., Chang Y. Feasibility study of RFID-Mounted drone application in management of oyster farms // International Geoscience and Remote Sensing Symposium (IGARSS). — 2017. — T. 2017—July. — C. 3610—3613.
4. Almalki F. A. Utilizing drone for food quality and safety detection using wireless sensors // 2020 3rd IEEE International Conference on Information Communication and Signal Processing, ICICSP 2020. — 2020. — C. 405—412
5. M. Longhi [и др.] RFIDrone: Preliminary experiments and electromagnetic models 2016 URSI International Symposium on Electromagnetic Theory, EMTS 2016. C. 450—453.
6. The Interrogation Footprint of RFID-UAV: Electromagnetic Modeling and Experimentations / G. Casati [и др.] // IEEE Journal of Radio Frequency Identification. — 2017. — T. 1, No 2. — C. 155—162. — DOI: 10.1109/JRFID.2017.2765619.
7. A New Vision for Smart Objects and the Internet of Things: Mobile Robots and Long-Range UHF RFID Sensor Tags / J. Wang [и др.]. — 2015. — Июль. — arXiv: 1507.02373. — URL: <http://arxiv.org/abs/1507.02373>.

8. RFID and Drones: The Next Generation of Plant Inventory / J. Quino [и др.] // AgriEngineering. — 2021. — Т. 3, № 2. — С. 168—181. — DOI: 10.3390/agriengineering3020011.
9. Buffi A., Nepa P., Cioni R. SARFID on drone: Drone-based UHF-RFID tag localization // 2017 IEEE International Conference on RFID Technology and Application, RFID-TA 2017. — 2017. — С. 40—44. — DOI: 10.1109/RFID-TA.2017.8098872.
10. A SAR-Based Measurement Method for Passive-Tag Positioning with a Flying UHF-RFID Reader / A. Buffi [и др.] // IEEE Transactions on Instrumentation and Measurement. — 2019. — Т. 68, № 3. — С. 845—853. — DOI: 10.1109/TIM.2018.2857045.
11. Vishnevskiy V., Larionov A., Ivanov R. Analysis and simulation of UHF RFID vehicle identification system // Communications in Computer and Information Science. — 2016. — Т. 678. — С. 35—46. — DOI: 10.1007/978-3-319-51917-3-4.

UDC: 004.7

Collision Provenance using Decentralized Ledger as a Blockchain/Hashgraph in Swarm of Drones

Q. Zirak¹ and D.V. Shashev¹

¹National Research Tomsk State University, Russian Federation, Tomsk

qazawatzirak@gmail.com, dshashev@mail.ru

Abstract

There has been a lot of research on drone swarms in recent years mostly focusing on the swarm pattern formation and communication aspects improving the coverage and latency of communication, however; the decentralization aspect of the swarm network is little in consideration. A peer-to-peer flying ad-hoc network might be a good solution to implement a decentralized swarm, but besides the physical layer of the network, there must be a mechanism which ensures the preservation, validity, immutability, and integrity of data ever emitted in the swarm. Since a decentralized network is not a leader-based network, it needs a mechanism for all the participants in the network to reach consensus on all actions. This is where a shared data structure in the form of blockchain/hashgraph can be introduced in application layer. A change in state requires a consensus by all participants and mirroring of state change. We can leverage immutable smart contracts in such a way that drones upload data to the topics created in those networks. Since the data can not be altered, it creates a trust layer for the drones, hence a good approach for provenance of collisions.

Keywords: Blockchain, hashgraph, decentralization, consensus, collision, Ethereum, Hedera

1. Introduction

Decentralization begins in the physical layer of the network in the form of flying ad-hoc network. All the intermediate drones from a source to destination begin routing of packets as compared to a centralized model which takes care of packet delivery in the form of client-server architecture. Proactive and reactive routing protocols get the best path for packet delivery. However, our motivation goes beyond this layer and is rooted in the application layer of the network. Combining a shared data structure that maintains the same state across all participants is a

Supported by Tomsk State University Development Program (Priority-2030).

good solution to overcome the problems with preservation, validity and integrity of data. Furthermore, to achieve decentralization it is necessary that the data recorded is immutable. In other words, governance is required for alteration of recorded data. This is where we can integrate ledgers like blockchain or hashgraph.

There are a lot of Distributed Ledgers (DLTs) being developed and are still in development. Bitcoin [1], Ethereum [2], Hedera, Solana [3], Avalanche, Cosmos, Polkadot and more. These DLTs all have some kind of consensus mechanism such as Proof-of-Work, Proof-of-Stake, or Proof-of-History. The purpose of this consensus mechanism is to let the network decide which participant of the network gets to update the shared database. Each participant requires some form of account with private and public key, and this shared database is stored by each participant as a copy. It is to be noted that DLT network is maintained by an external set of nodes. Participants in swarm of drones do not store copy of the ledger but utilize a ‘greatly distributed’ DLT. Greatly distributed means that the nodes are not pooled together to do a 51 percent attack on the network[4]. All the previously mentioned DLTs are greatly distributed and hence can be used for provenance. The only part of the play for participants of the swarm is to periodically publish GPS information to topics created in a DLT. Hence, published GPS information from different manufacturers becomes immutable and reliable for provenance of collision.

Blockchain in its very basic form is like a singly linked list, which is a series of customized data types or objects. It is a series or chain of blocks of information [5]. All the blocks have information in them in the form of transactions. A transaction is an input which changes the state of the DLT. These transactions are hashed together to get a single hash value. This hash value is equivalent to the pointer or memory location in a linked list. When a new block is added to the chain, it stores this hash value of the previous block in itself. The hashing is not random but done based on a Merkle Tree. All the transactions are leaf nodes and the hash value which is stored in the next block as a reference to the previous one is root hash of the Merkle Tree of that block. This hash is powerful because if any single bit of information in a block changes or is maliciously being changed, the whole hash calculated for that block changes. If previous block’s hash is changed, the next block which had stored its hash already will be invalid. This creates immutability of information.

Hashgraph is a DLT that utilizes Directed Acyclic Graph (DAG) instead of blockchain and it can be deemed as the next level of blockchain as it introduces blazing fast transactions speed and virtual voting[6]. Each participant in the network stores a copy of the ledger. As time passes, each participant shares state of the ledger with random other participants. Those participants which received this information create an event. This event includes a hash of the event which it received from a participant and hash of the event of its own previous event. This is continued

throughout the network. This is the standard Gossip protocol. But not only that, in the Hashgraph consensus[7], Gossips about Gossip are being propagated throughout the network resulting in exponential spread of information. Unlike blockchains where you have a block creation time, in Hash-graph consensus, there is no need for block formation time or artificial delay between blocks. As for governance of the network, voting is required. But, since every other participant knows what every other participant has stored because of Gossip protocol, the voting occurs virtually. There is no need for transactions for voting because every participant can predict what every other participant is going to vote for.

2. Architecture

2.1. Network of DLT. Provenance of collision requires that the DLT network should be greatly distributed. It can be built up from the scratch specialized for swarm of drones but that diminishes the purpose of decentralization[8]. A ‘shared airspace’ between different manufacturers requires a DLT that is not proprietary. In other words, swarm networks from different manufacturers leverage the trust of already available public DLTs. Two such greatly distributed DLTs are Ethereum and Hedera. Swarms from different manufacturers will be publishing GPS information periodically to the DLT network as shown in Fig.1.

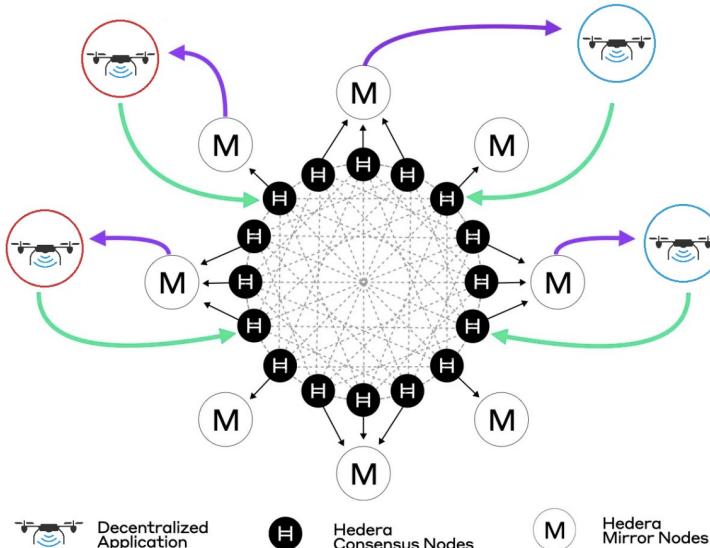


Fig.1. Hedera Network with Consensus and Mirror nodes. Drones in same colored circles are of same network and manufacturer publishing GPS data

Since, drones in swarms will be in fast-paced dynamic movement, the interval of publishing data to DLT should be carefully considered[9]. The reason is that every state change to the DLT via a transaction requires a cost in the form of native token of the DLT. This cost is calculated based on gas. Gas (EIP-1559) is the measurement unit that reflects the cost necessary to pay for computational resources used to process transactions[10]. For example in Ethereum, the update of Storage costs around 20,000 gas. Things can be further improved such as only publishing GPS data if there is another object in proximity.

2.2. Smart contracts. Smart contracts are small pieces of reactive programs that are deployed onto the DLT. They are digital contracts in nature and require a set of private and public key (Asymmetric encryption) to interact with them. They do not execute by themselves and require external agents to execute them. In our case of swarms, a small back-end program can trigger the functions of a smart contract. As mentioned previously, each line of a smart contract that changes the state of the DLT has a gas cost. The smart contracts are usually written in Solidity language, but the Ethereum Virtual Machine (EVM) can be directly programmed for gas optimization using the intermediate assembly language Yul.

Since Ethereum DLT has been adopted in large numbers, which made other DLTs to include compatibility of EVM. Hedera is one of the DLTs which is EVM compatible. In other words, smart contract written for EVM can be deployed directly onto Hedera Hashgraph. The gas cost of contract deployment and interaction on Hedera is quite cheap and transaction speed, because of adopting DAG, is quite fast. Gas cost of some of the opcodes can be seen in Table 1.

Operation	London(Gas)	Current(Gas)
SLOAD	2,100	2,100
SSTORE	22,100	Max
CALL	2,600	2,600
LOG0-4	375*2*topics + data Mem	Max

Table 1. Opcodes gas fees

3. Experiment

For actual implementation we have two ways to implement it. First is to create a single smart contract that is shared among all participants from different manufacturers. Second is to create a Factory Contract just like Uniswap's factory contract

that can be interacted with to create child GPS Storage contracts using the Create2 Opcode of EVM. Create2 opcode of EVM derives a predictable address of a smart contract to be deployed prior to its actual deployment. But the problem is, Hedera does not yet support Create2 opcode of EVM. Therefore, a good approach is to create a Factory Contract that uses Create opcode instead of Create2.

The apparatus consists of:

- 1) Hedera SDK - Javascript Variant
- 2) Hedera Account
- 3) Smart Contracts (Factory, Child GPS)
- 4) Back-end Services for automation

The same set of Smart contracts can also be deployed to Ethereum and will require:

- 1) Ethereum Account
- 2) Infura or Moralis RPC and project ID

For Hedera Network we need to use Hedera SDK to deploy the smart contracts:

- 1) Upload ByteCode to Hedera Network using *FileCreateTransaction()*
- 2) Instantiate the Smart Contract byte code using *ContractCreateTransaction()*
- 3) Interact with the deployed contract using *ContractExecuteTransaction()*

3.1. Factory Contract. The purpose of Factory Contract is nothing more than to create Child GPS Contracts and keep reference to them. There is only one Factory Contract without any Owners to keep it completely decentralized. The following functions must be implemented in the Factory Contract:

- 1) Function *createTracking(droneID)* - Deploys/Overrides a new Child GPS Contract for the 'droneID'. 'droneID' is public address which the drone will be using to send transactions. The 'msg.sender' or caller of the function must be the 'droneID'.
- 2) Event *TrackingCreated(droneID indexed)* - An indexed event emitted when tracking is created.

3.2. Child GPS Contract. Everything related to tracking GPS coordinates and usage logic should be placed in this contract for each individual drone. For tracking history of variables, we can either create our own data structure inside Child GPS Contract or rely on transactions history of DLT of specified Child GPS Contract. But since, we need to know history of variables not just transaction history, we will be creating our own data structure. Each drone can have one Child GPS Contract with the following base of functions:

- 1) Function *publish(coordinates)* - Updates state of DLT with GPS coordinates supplied as a struct parameter to the function. Coordinates struct can have required variables such as Altitude, Latitude, Longitude, Speed, Direction and more. The caller of the function must be the owner of the Child GPS Contract.
- 2) Function *alertProximity()* - This function is called when proximity of another object is detected by the back-end service. This in turn calls *publish(coordinates)* and emits an event of alert. The reason of creating this function is to easily track collision history by searching for *Proximity(coordinates)* event. Caller must be the owner of the contract. Details of back-end service is out of the scope of this paper.
- 3) Event *Published(coordinates indexed)* - An event emitted when coordinates are published.
- 4) Event *Proximity(coordinates indexed)* - An event emitted when a proximity is alerted.

4. Conclusion

Combining traditional ad-hoc network in swarm of drones with an application layer trust in the form of a distributed ledger creates a powerful history of tracking for swarms in a shared airspace. The data recorded is averaged in Table 2.

Finality (Avg)	publish (Gas)	alertProximity(Gas)	createTracking (Gas)
4.5 seconds	104,075	106,273	1,228,565

Table 2. Gas costs of functions

Keeping in mind that Hedera charges 0.0000000852 dollars per unit of gas, the actual gas costs can be seen in Table 3.

publish	alertProximity	createTracking
0.00886719 \$	0.0090544596 \$	0.104673738 \$

Table 3. Gas costs of functions in dollars

As it can be seen that if GPS coordinates are published every 2 seconds, total cost of a day would be 400\$ approximately. This can be greatly reduced by increasing interval or if the publishing is only done when in proximity of another object or more specifically drone. Hence, if a collision happens between two drones from different manufacturers, it is not possible to alter collision data by these manufacturers. Thus, a reliable source for provenance of collision.

DLTs are still in their very early stage of adoption and development. The finality times and gas costs of DLTs for operation in the Internet of Things are still costly however, this is an active field of research. When it comes to Swarm of Drones, there have been quite few works to integrate a DLT with swarm. This integration greatly benefits and adds to decentralization aspect. Thus, this paper shows a basic practical application of DLTs in swarm of drones.

REFERENCES

1. Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System". 2008.
2. Ethereum white paper, <https://ethereum.org/en/whitepaper/>. 2014.
3. Anatoly Yakovenko, "Solana: A new architecture for a high performance blockchain v0.8.13". 2017
4. Sayeed, Sarwar & Marco-Gisbert, Hector. (2019). Assessing Blockchain Consensus and Security Mechanisms against the 51% Attack. Applied Sciences. 9. 1788. 10.3390/app9091788.
5. I. J. Jensen, D. F. Selvaraj and P. Ranganathan, "Blockchain Technology for Networked Swarms of Unmanned Aerial Vehicles (UAVs)," 2019 IEEE 20th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM), 2019, pp. 1-7, doi: 10.1109/WoWMoM.2019.8793027.
6. Dr. Leemon Baird, Mance Harmon, Paul Madsen, "Hedera: A Public Hashgraph Network & Governing Council", August 2020.
7. Diving deep into Hashgraph consensus, <https://hedera.com/learning/what-is-hashgraph-consensus>
8. Alsamhi, Saeed & Lee, Brian & Guizani, Mohsen & Kumar, Neeraj & Qiao, Yuansong & Liu, Xuan & Alsamhi, S. (2021). Blockchain for decentralized multi-drone to combat COVID-19 and future pandemics: Framework and proposed solutions. Transactions on Emerging Telecommunications Technologies. 32. 10.1002/ett.4255.
9. Tonghe Wang, Songpu Ai, and Junwei Cao. "A Blockchain-Based Distributed Computational Resource Trading Strategy for Industrial Internet of Things Considering Multiple Preferences ". 2022
10. Yulin Liu, Yuxuan Lu et.al. "Empirical Analysis of EIP-1559: Transaction Fees, Waiting Time, and Consensus Security". 2022

UDC: 519.23

Retrial Queuing System with Limited Processor Sharing Discipline

A.N. Dudin^{1,2}, S.A. Dudin¹, O.S. Dudina¹

¹Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus

²Department of Applied Probability and Informatics, RUDN University, 6, Miklukho-Maklaya st., 117198, Moscow, Russia
dudin@bsu.by, dudins@bsu.by, dudina@bsu.by

Abstract

We consider a retrial queueing system with limited processor sharing discipline. The arrival flow is defined by a Markovian arrival process. The distinguishing feature of the model is that the customers on service can interfere each other and the effective bandwidth of the server decreases with the increase in the number of servicing customers. The process of the system states is defined as level-dependent Markov chain. The generator of this chain is derived. The main performance measures of the system are obtained.

Keywords: Markovian arrival flow, processor sharing, retrials

1. Introduction

Queuing systems are effectively used to model and optimize various industrial, logistics and telecommunications systems and networks. In some of these systems, customers are serviced one at a time in the order specified by the service discipline. However, often, customers can be serviced in the system at the same time. In this case, multi-server systems are considered. That is, the system throughput is divided into several parts, called as servers, and each server can serve one customer. Multi-server queuing systems are a popular subject for research. An overview of the state of the art can be found, for example, in [1]. It should be noted that multi-server systems have their drawbacks in terms of optimal use of the system resource. For example, in a situation where there is one customer for service, and there are many servers, then the bulk of the bandwidth is not used. As an alternative to multi-server systems, queuing systems with the discipline of processor sharing are considered. For a review of work on processor sharing systems, see e.g. [2, 3, 4]. This discipline assumes that the entire resource of the system provides simultaneous

service to all customers available for servicing. That is, even when one customer is being serviced, the system resource is fully used.

This work is devoted to the study of a queuing system with the discipline of processor sharing. Compared with classical systems, this model has the following features that increase its adequacy to modern systems. First, we assume that the customer has a required service rate that cannot be exceeded. In fact, if a user of a wireless communication network requires a certain amount of system bandwidth to operate, then it is not reasonable to allocate all the system bandwidth to the user. He/she simply would not be able to use it and would not be serviced faster. However, if there are a lot of customers in service, and there is not enough bandwidth to provide service to all customers at the required rate, then a decrease in the average service rate is allowed. Second, we assume that the number of customers in service is limited by a given control parameter. If we do not restrict access to the system, then a situation may arise when the number of customers in service will be so large that customers will be serviced at an unacceptably low service rate. Third, we assume that customers in service may interfere with each other. In fact, in wireless communication networks, in order to avoid interference and organize multiple access, part of the system bandwidth can be spent on delimiting requests from users. In addition, in this paper, we assume that the input flow of customers is specified by a *MAP* (Markovian arrival process), see, for example, [5], which allows significant traffic fluctuations inherent in modern telecommunications communication networks. Also, for greater model adequacy, we assume that customers that were not allowed to be serviced upon arrival in the system can make repeated attempts to get serviced. In other words, we consider a system with repeated calls. The current state of the problem of studying the retrial systems is described in [6].

2. Mathematical model

We consider a retrial queuing system with a limited processor sharing service discipline.

The structure of the system is shown in Figure 1.

A single server can serve up to N customers at the same time. The buffer is missing. We assume that the total throughput of the server is equal to M megabits per second. One customer requires an average speed of X megabits per second to serve. The average volume of one customer is S megabits. Thus, if a customer is serviced with the required bandwidth, then its average service time is defined as $b_1 = S/X$. In this paper, we will assume that the service time of one customer has an exponential distribution. If the required bandwidth is available, the exponential service time distribution parameter is given as $\mu = \frac{1}{b_1}$. We suggest that the serviced customers can interfere with each other, that is, if there are n , $n = \overline{2, N}$, customers

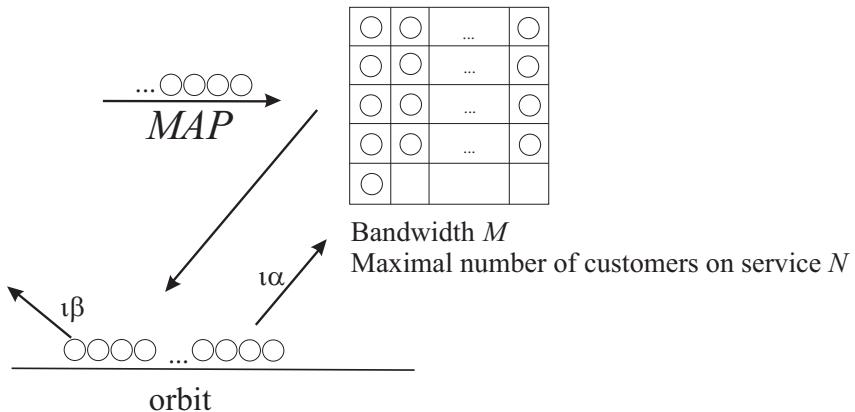


Fig. 1. Structure of the system

in service, then the effective throughput of the server is M_n , and, $M_n \geq M_{n+1} > 0$ for any admissible n . If there are such a number of customers n in service that $nX \leq M_n$, that all customers receive the required service rate and are served with intensity μ . Otherwise, each customer is allocated bandwidth $X_n = \frac{M_n}{n}$ megabits per second and its service rate is $\mu_n = \frac{X_n}{S}$.

The *MAP* arrival process enters the system. This flow is specified by the underlying process ν_t , $t \geq 0$, which is an irreducible Markov chain with continuous time and finite state space $\{1, 2, \dots, W\}$, and matrices D_0 and D_1 . Let us denote the average intensity of incoming customers as λ . A detailed description of the *MAP* process and its properties, as well as formulas for its characteristics can be found in [1].

If at the moment of arrival of a customer the number of customers receiving service is less than the parameter N , then the customer is admitted for service. Otherwise, the customer goes to an orbit of unlimited capacity, from where it makes repeated attempts to get serviced at exponentially distributed intervals with the parameter α . An attempt is considered successful if, at the time of its execution, the number of service customers is less than N . In this case, the customer starts service, and the number of customers in the orbit reduces by one. If the attempt was unsuccessful, that is, at the time of its execution, the number of service customers was equal to N , then the customer returns to orbit. The customers staying in orbit can be impatient. This means that each customer from the orbit can leave it after an exponentially distributed time with the parameter β , $\beta > 0$.

We assume that integer N is the control parameter and the purpose of analytical modeling is to determine the optimal value N , at which the probability of losing an

arbitrary customer due to impatience admits the minimal value for any fixed set of the system parameters.

3. The process of system states and its analysis

The behavior of the system under consideration can be described by the regular irreducible continuous-time Markov chain

$$\xi_t = \{i_t, n_t, \nu_t\}, t \geq 0,$$

where at time t , $t \geq 0$, i_t is the number of customers in the orbit, $i_t \geq 0$; n_t is the number of customers on service, $n_t = \overline{0, N}$; ν_t is the state of the underlying process of the MAP, $\nu_t = \overline{1, W}$.

Theorem 1. The infinitesimal generator Q of Markov chains ξ_t , $t \geq 0$, has a block tridiagonal structure

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & O & \dots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & O & \dots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & O & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where the non-zero blocks $Q_{i,j}$, $|i - j| \leq 1$, are defined as follows:

$$Q_{i,i} = \begin{pmatrix} Q_{i,i}^{(0,0)} & Q_{i,i}^{(0,1)} & O & \dots & O & O \\ Q_{i,i}^{(1,0)} & Q_{i,i}^{(1,1)} & Q_{i,i}^{(1,2)} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & Q_{i,i}^{(N,N-1)} & Q_{i,i}^{(N,N)} \end{pmatrix},$$

$$Q_{i,i}^{(0,0)} = D_0 - i(\alpha + \beta)I_W,$$

$$Q_{i,i}^{(n,n)} = D_0 - i(\alpha + \beta)I_W - n\mu I_W, n \leq \frac{M_n}{X},$$

$$Q_{i,i}^{(n,n)} = D_0 - i(\alpha + \beta)I_W - n\mu_n I_W, \frac{M_n}{X} < n < N,$$

$$Q_{i,i}^{(N,N)} = D_0 - i\beta I_W - N\mu_N I_W,$$

$$Q_{i,i}^{(n,n+1)} = D_1,$$

$$Q_{i,i}^{(n,n-1)} = n\mu I_W, n \leq \frac{M_n}{X}, Q_{i,i}^{(n,n-1)} = n\mu_n I_W, \frac{M_n}{X} < n \leq N,$$

$$Q_{i,i-1} = \begin{pmatrix} Q_{i,i-1}^{(0,0)} & Q_{i,i-1}^{(0,1)} & O & \dots & O & O \\ O & Q_{i,i-1}^{(1,1)} & Q_{i,i-1}^{(1,2)} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & Q_{i,i-1}^{(N-1,N-1)} & Q_{i,i-1}^{(N-1,N)} \\ O & O & O & \dots & O & Q_{i,i-1}^{(N,N)} \end{pmatrix},$$

$$Q_{i,i-1}^{(n,n)} = i\beta I_W, \quad 0 \leq n \leq N, \quad Q_{i,i+1}^{(n,n+1)} = i\alpha I_W, \quad 0 \leq n \leq N-1,$$

$$Q_{i,i+1} = \begin{pmatrix} O & O & \dots & O & O \\ O & O & \dots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \dots & O & O \\ O & O & \dots & O & D_1 \end{pmatrix}.$$

The proof of the theorem is carried out by carefully analyzing all possible transitions of the Markov chain ξ_t and further grouping the intensities into generator blocks.

Note that due to the assumption that customers in the orbit are impatient, it is easy to show that a stationary distribution of the system states exists for any values of the system parameters.

Denote by $\pi(i, n, \nu)$, $i \geq 0$, $n = \overline{0, N}$, $\nu_t = \overline{1, W}$, the stationary probabilities of the states of the chain ξ_t . Let us form the row vectors from these probabilities

$$\boldsymbol{\pi}(i, n) = (\pi(i, n, 1), \dots, \pi(i, n, W)), \quad i \geq 0, \quad n = \overline{0, N},$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i, 0), \dots, \boldsymbol{\pi}(i, N)), \quad i \geq 0.$$

To find the vectors of stationary probabilities $\boldsymbol{\pi}_i$, $i \geq 0$, it is recommended to use the efficient algorithm developed in [7].

4. Performance measures

The mean number of customers on service is $N_{serv} = \sum_{i=0}^{\infty} \sum_{n=1}^N n \boldsymbol{\pi}(i, n) \mathbf{e}$.

The mean number of customers in the orbit is $N_{orbit} = \sum_{i=1}^{\infty} i \boldsymbol{\pi}_i \mathbf{e}$.

The mean number of customers in the system is $L = \sum_{i=0}^{\infty} \sum_{n=0}^N (i + n) \boldsymbol{\pi}(i, n) \mathbf{e} = N_{serv} + N_{orbit}$.

The probability that the system is idle at arbitrary moment is calculated as $P_{idle} = \boldsymbol{\pi}(0, 0) \mathbf{e}$.

The intensity of the flow of served customers is computed as

$$\lambda_{out} = \sum_{i=0}^{\infty} \sum_{n=1}^N \left(\delta_{n \leq \frac{M_n}{X}} n \mu \pi(i, n) \mathbf{e} + \delta_{\frac{M_n}{X} < n < N} n \mu_n \pi(i, n) \mathbf{e} \right),$$

where $\delta_a = \begin{cases} 1, & \text{if } a \text{ is true,} \\ 0, & \text{otherwise.} \end{cases}$

The probability that an arbitrary customer will be lost is calculated as

$$P_{loss} = \frac{1}{\lambda} \sum_{i=1}^{\infty} i \beta \pi_i = 1 - \frac{\lambda_{out}}{\lambda}.$$

The probability that, at an arbitrary moment, customers receive a reduced service rate is computed as $P_{sharing} = \sum_{i=0}^{\infty} \sum_{n=1}^N \delta_{\frac{M_n}{X} < n < N} \pi(i, n) \mathbf{e}$.

5. Conclusion

In this paper, we consider a retrial queueing model with processor sharing discipline. The results can be expanded to the case of customers impatience during service in the case of shortage of the desirable bitrate due to sharing.

REFERENCES

1. Dudin A., Klimenok V. I., Vishnevsky V. M. The Theory of Queueing Systems with Correlated Flows. Cham: Springer. 2020. P. 1-410.
2. Yashkov S. F., Yashkova A. S. Processor sharing: A survey of the mathematical theory // Automation and Remote Control. 2007. V. 68(9). P. 1662-1731.
3. Altman E., Avrachenkov K., Ayesta U. A survey on discriminatory processor sharing // Queueing systems. 2006. V. 53. N. 1. P. 53-63.
4. Kim C., Dudin S. A., Dudina O. S., Dudin A. N. Mathematical models for the operation of a cell with bandwidth sharing and moving users // IEEE Transactions on Wireless Communications. 2019. V. 19(2). P. 744-755.
5. Chakravarthy S. R. The batch Markovian arrival process: a review and future work, in: A. Krishnamoorthy, N. Raju, V. Ramaswami (Eds.), Advances in Probability Theory and Stochastic Processes, Notable Publications Inc., New Jersey. 2001. P. 21-29.
6. Kim J., Kim B. A survey of retrial queueing systems // Annals of operations research. 2016. V. 247. N. 1. P. 3-36.
7. Dudin S., Dudina O. Retrial multi-server queueing system with PHF service time distribution as a model of a channel with unreliable transmission of information // Applied Mathematical Modelling. 2019. V. 65. P. 676-695.

УДК: 519.23

О распределении числа подряд потерянных запросов в системе $MAP/PH/1/N$

В.И. Клименок, А.Н. Дудин

Факультет прикладной математики и информатики
Белорусский государственный университет
проспект Независимости, 4, Минск, Беларусь
klimenok@bsu.by, dudin@bsu.by

Аннотация

В статье предлагаются методы вычисления важной вероятностной характеристики производительности системы массового обслуживания с конечным буфером – распределения числа подряд потерянных запросов. Рассматривается система $MAP/PH/1/N$. Это однолинейная система с конечным буфером размера N , марковским потоком (общепризнанная аббревиатура MAP - Markovian Arrival Process) и фазовым распределением времен обслуживания (PH - Phase Type distribution). Наиболее известной и важной оценкой производительности такой системы является вероятность потери произвольного запроса. Вероятность потери является предметом исследования в литературе при различных предположениях о характере входного потока и распределении времен обслуживания. В то же время эта характеристика не всегда может быть хорошей оценкой качества обслуживания в системах массового обслуживания, возникающих при математическом моделировании телекоммуникационных сетей. Более показательной в этом случае является вероятность потери нескольких запросов подряд. В данной статье мы предлагаем явные формулы для вычисления распределения числа подряд потерянных запросов с момента, когда в буфере остается одно свободное место, и соответствующего математического ожидания.

Ключевые слова: марковский входной поток, конечный буфер, потеря серии запросов, математическое ожидание серии

1. Введение

Системы массового обслуживания с потерями образуют практически важный и интересный в математическом плане класс систем в теории массового обслуживания. Обзор ранних работ по таким системам можно найти в книге [1].

Значительное число результатов в этой области получены в Российском университете дружбы народов, см., например, книгу [2] и многочисленные статьи Бочарова П.П. с соавторами. В последние десятилетия в связи с бурным развитием телекоммуникационных сетей и усложнением характера трафика в таких сетях широкое применение получили такие математические модели входных потоков как *MAP* и *BMAP* (Batch Markovian Arrival Process), которые хорошо описывают коррелированный взрывной трафик и, вместе с тем, допускают прозрачную математическую трактовку. Системы с *MAP* и *BMAP* потоками и потерями рассмотрены в ряде статей при различных предположениях о дисциплине приятия в систему и распределении времен обслуживания, см., например, статьи [3, 4, 5] и ссылки в них.

Наиболее известной и важной оценкой производительности систем с потерями, обусловленными конечным размером буфера, является вероятность потери (отказа в обслуживании) произвольного запроса, поступившего с систему и заставшего буфер полностью занятым. Большинство исследователей при оценке потерь в таких системах ограничиваются вычислением именно этой вероятности. Однако, как подчеркнуто в [6], только вероятность отказа в обслуживании не может вполне характеризовать качество обслуживания в телекоммуникационных сетях. Более эффективной оценкой потерь является вероятность потери нескольких запросов подряд. Вместе с тем, как показано в [7], вероятность потери нескольких запросов подряд нельзя вывести из вероятности отказа в обслуживании. Известен ряд работ, где рассматриваются задачи вычисления вероятности потери более чем k запросов подряд за период занятости системы (в англоязычной литературе $k - CCL$ – k -consecutive customer loss). Такие задачи для систем $M/G/1/n$, $GI/M/1/n$ и их групповых аналогов рассматривались в работах [8, 9, 10, 11] и ряда других авторов и исследовались путем составления и решения системы линейных алгебраических уравнений (СЛАУ) для вероятностей того, что за период занятости, порожденный i запросами в системе, не произойдет $k - CCL$ событие. Используя идею, развитую в процитированных работах, авторы [5] составили аналогичную СЛАУ для систем массового обслуживания $BMAP/G(a; b)/1/N$ и $BMAP/MSP(a; b)/1/N$. В данном случае СЛАУ имеет сложный вид, поскольку коэффициенты представлены матрицами, вычисление которых является самостоятельной задачей. Вместе с тем, вопросы решения полученной СЛАУ не обсуждаются. Важным для приложений является среднее число подряд пропущенных запросов, известное в телекоммуникационной литературе как average packet gap (среднее число пропущенных пакетов) широко используется при исследовании телекоммуникационных сетей, см., например, [12].

В данной статье мы предлагаем явные формулы для вычисления распределения числа подряд потерянных запросов с момента, когда в буфере остается одно свободное место, и соответствующего математического ожидания для системы $MAP/PH/1/N$.

2. Вычисление стационарного распределения и среднего числа подряд потерянных запросов в системе $MAP/PH/1/N$

Запросы поступают в рассматриваемую систему в MAP -потоке под управлением неприводимой цепи Маркова с непрерывным временем ν_t , $t \geq 0$, которая принимает значения в множестве $\{0, 1, 2, \dots, W\}$ и называется управляющим процессом MAP . Интенсивности переходов управляющего процесса, сопровождающиеся генерацией запроса, задаются $(W+1) \times (W+1)$ матрицей D_1 , а переходы, не сопровождающиеся генерацией запроса – недиагональными элементами $(W+1) \times (W+1)$ матрицы D_0 . Диагональные элементы матрицы D_0 есть взятые со знаком минус интенсивности выхода управляющего процесса MAP из соответствующих состояний. Более подробную информацию о MAP с указанием первоисточников можно найти в [13]. Далее нам понадобятся обозначения и минимальные сведения о некоторых характеристиках MAP :

- $\boldsymbol{\theta}$ – вектор-строка стационарных вероятностей состояний MAP , он вычисляется как единственное решение СЛАУ $\boldsymbol{\theta}D(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$;

- $P(k, t)$ - квадратная матрица порядка $W + 1$, у которой (ν, ν') -й элемент есть вероятность того, что в интервале длительностью t поступит k запросов и управляющий процесс MAP перейдет в состояние ν' при условии, что в начале этого интервала он находился в состоянии ν . Производящая функция этих матриц имеет вид: $P(z) = \sum_{k=0}^{\infty} P(k, t)z^k = e^{D(z)t}$, где $D(z) = D_0 + D_1z$.

Время обслуживания запроса имеет PH распределение, заданное неприводимым представлением $(\boldsymbol{\beta}, S)$ и управляющим процессом (цепью Маркова) m_t , $t \geq 0$, с пространством состояний $\{1, \dots, M, M + 1\}$, где состояние $M + 1$ является поглощающим. В начале обслуживания состояние (фаза обслуживания) выбирается в соответствии с вектором- строкой $\boldsymbol{\beta}$. Далее начинается блуждание в пространстве несущественных состояний в соответствии с $M \times M$ матрицей интенсивностей переходов S . Время задержки в состоянии t имеет показательное распределение с параметром $(-S)_{m,m}$. Вероятность перехода управляющего процесса из состояния t в состояние t' за время t определяется как $(e^{St})_{m,m'}$. После попадание в поглощающее состояние обслуживание запроса заканчивается. Интенсивности переходов в поглощающее состояние определяются вектором- столбцом $\mathbf{S}_0 = -S\mathbf{e}$. Более подробное описание PH распределения с указанием первоисточников можно найти в [13].

Теперь мы можем перейти к нахождению искомого распределения числа подряд потерянных запросов с момента, когда в буфере остается одно свободное место в системе $MAP/PH/1/N$. Пусть n_t - число запросов в буфере; ν_t и m_t - состояния управляющих процессов MAP и PH соответственно в момент времени t . Предполагаем, что в конце периода занятости всегда устанавливается фаза обслуживания для запроса, инициирующего следующий период занятости. Тогда функционирование системы описывается цепью Маркова $\xi_t = \{n_t, \nu_t, m_t\}, t \geq 0$, Обозначим через π_n вектор стационарных вероятностей состояний процесса $\xi_t, t \geq 0$, соответствующих n запросам в буфере, $n = \overline{1, N}$. Предполагаем, что компоненты этого вектора упорядочены в лексикографическом порядке. Алгоритм вычисления векторов $\pi_n, n = \overline{1, N}$, хорошо известен, см., например, [2], [14]. В данной статье мы не приводим этот алгоритм из соображений краткости изложения.

Обозначим через q_k вероятность того, что в произвольный момент времени в буфере есть одно свободное место и с этого момента потеряются k запросов подряд. Анализируя ситуации, связанные с потерей k событий подряд, можно увидеть, что такие потери возможны тогда и только тогда, когда в буфере остается одно свободное место, поступает запрос, который занимает это место, и за время дообслуживания текущего запроса (за время обслуживания, если $N = 1$) в MAP поступит k запросов. Из сказанного следует, что

$$q_k = \frac{\pi_{N-1}(D_1 \otimes I_M)}{\lambda} \int_0^{\infty} P(k, t) \otimes e^{St} dt (\mathbf{e}_{W+1} \otimes \mathbf{S}_0). \quad (1)$$

Основная проблема при вычислении вероятностей q_k заключается в вычислении интегралов $Y_k = \int_0^{\infty} P(k, t) \otimes e^{St} dt$. Мы можем предложить следующие способы вычисления этих интегралов.

1). Метод униформизации матричной экспоненты, см. [13], стр.110-111. Используя этот метод, получим следующее разложение для матрицы $P(k, t)$:

$$P(k, t) = \sum_{j=0}^{\infty} e^{-\tilde{\theta}t} \frac{(\tilde{\theta}t)^j}{j!} K_k^{(j)}, \quad k \geq 1, \quad (2)$$

где $\tilde{\theta} = \max_{i=\overline{0, W}} (-D_0)_{ii}$, а матрицы $K_k^{(j)}$ удовлетворяют следующей системе рекуррентных соотношений:

$$K_0^{(0)} = I, \quad K_k^{(0)} = O, \quad k \geq 1, \quad K_k^{(j+1)} = \tilde{\theta}^{-1} K_{k-1}^{(j)} D_1 + K_k^{(j)} (I + \tilde{\theta}^{-1} D_0), \quad k \geq 0, \quad j \geq k.$$

Используя разложение (2), интегралы Y_k можно вычислить в явном виде и для вероятностей (1) получаются явные формулы.

2). В этом методе используется рекуррентное вычисление интегралов Y_k , аналогичное описанному в [13], стр.125.

Рекуррентная процедура строится следующим образом:

$$Y_0 = -(D_0 \oplus S)^{-1}, \quad Y_k = \sum_{i=0}^{k-1} Y_i (D_{k-i} \otimes I_M) (D_0 \oplus S)^{-1}, k \geq 1.$$

3). Теоретически для вычисления интегралов Y_k можно использовать метод производящих функций. Известно, что производящая функция матриц $P(k, t)$ имеет вид $P(z, t) = \sum_{k=0}^{\infty} P(k, t) z^k = e^{D(z)t}$. Тогда производящая функция вероятностей q_k вычисляется как

$$Q(z) = \sum_{k=0}^{\infty} q_k z^k = \frac{\pi_{N-1}(D_1 \otimes I_M)}{\lambda} (D(z) \oplus S)^{-1} (\mathbf{e}_{W+1} \otimes \mathbf{S}_0). \quad (3)$$

Вычислив вероятности $q_k, k \geq 1$, можно вычислить среднее число подряд потерянных запросов после попадания буфера в состояние $N - 1$. Это среднее можно вычислять по определению как $\sum_{k=1}^{\infty} k q_k$. Однако, как следует из вышеизложенного, это требует больших вычислительных затрат. Вместе с тем, как было отмечено выше, это среднее может быть полезным при исследовании телекоммуникационных сетей. Поэтому представляется весьма важным получить явные формулу для этой характеристики. Нам удалось решить эту проблему. Результат представлен в следующей теореме.

Теорема 1. Среднее значение γ числа подряд потерянных запросов в системе MAP/PH/1/N вычисляется по следующей формуле:

$$\begin{aligned} \gamma = & -\frac{\pi_{N-1}(D_1 \otimes I_M)}{\lambda} \left\{ I_{W+1} \otimes S^{-1} + \right. \\ & \left. [(\mathbf{e}\theta - D(1))^{-1} \otimes I_M] [(D(1) \oplus S)^{-1} S + D(1) \otimes S^{-1} - I] \right\} (D_1 \otimes I_M) \mathbf{e}. \end{aligned} \quad (4)$$

Доказательство. Искомое среднее γ будем вычислять как производную производящей функции (3) по z в точке $z = 1$. Чтобы найти эту производную, достаточно продифференцировать в точке $z = 1$ выражение $(D(z) \oplus S)^{-1} (\mathbf{e}_{W+1} \otimes \mathbf{S}_0) = \int_0^{\infty} (e^{D(z)t} \otimes e^{St}) dt (\mathbf{e}_{W+1} \otimes \mathbf{S}_0)$ в (3). Соответствующие выкладки представим в виде

цепочки преобразований, где будем по ходу указывать источники, откуда следует правомерность нетривиальных переходов.

$$\begin{aligned}
 & \int_0^\infty (e^{D(z)t})'|_{z=1} \otimes e^{St} dt (\mathbf{e}_{W+1} \otimes \mathbf{S}_0) = \\
 & = \int_0^\infty \sum_{k=1}^\infty \frac{t^k}{k!} \sum_{l=0}^{k-1} D(1)^{k-l-1} D'(1) D(1)^l \otimes e^{St} dt (\mathbf{e}_{W+1} \otimes \mathbf{S}_0) = \{D(1)\mathbf{e} = \mathbf{0}^T\} = \\
 & = \int_0^\infty \sum_{k=1}^\infty \frac{t^k}{k!} D(1)^{k-1} D'(1) \otimes e^{St} dt (\mathbf{e}_{W+1} \otimes \mathbf{S}_0) = \{[13], c.113\} = \\
 & = \left\{ \int_0^\infty t I_{W+1} \otimes e^{St} dt + \sum_{k=2}^\infty \int_0^\infty \frac{t^k}{k!} [-(\mathbf{e}\boldsymbol{\theta} - D(1))^{-1} D(1)^k \otimes e^{St} dt] \right\} (D'(1) \otimes (-S))\mathbf{e} = \\
 & = \left\{ I_{W+1} \otimes S^{-2} - [(\mathbf{e}\boldsymbol{\theta} - D(1))^{-1} \otimes I_M] \int_0^\infty \sum_{k=2}^\infty \frac{t^k}{k!} D(1)^k \otimes e^{St} dt \right\} (D'(1) \otimes (-S))\mathbf{e} = \\
 & = \left\{ I_{W+1} \otimes S^{-2} - [(\mathbf{e}\boldsymbol{\theta} - D(1))^{-1} \otimes I_M] \int_0^\infty (e^{D(1)t} - t D(1) - I) \otimes e^{St} dt \right\} (D'(1) \otimes (-S))\mathbf{e} = \\
 & = - \left\{ I_{W+1} \otimes S^{-1} + [(\mathbf{e}\boldsymbol{\theta} - D(1))^{-1} \otimes I_M] [(D(1) \oplus S)^{-1} S + D(1) \otimes S^{-1} - I] \right\} (D_1 \otimes I_M)\mathbf{e}.
 \end{aligned}$$

Подставляя последнее выражение цепочки в (3), получаем искомое выражение (4). ■

3. Заключение

В данной статье получены формулы для вычисления распределения числа подряд потерянных запросов с момента, когда в буфере остается одно свободное место, и соответствующего математического ожидания в системе *MAP /PH/1/N*. Эти формулы могут служить для адекватной оценки качества обслуживания в телекоммуникационных сетях различного назначения. В продолжение работы планируется расширить результаты на случай *BMAP*, т.е., на системы с групповым марковским потоком.

Литература

1. Chaudhry M. L., Templeton J.G.C. First course in bulk queues. John Wiley & Sons, 1983.
2. Bocharov P. P., D'Apice C., Pechinkin A. V. Queueing Theory. Walter de Gruyter, 2011.
3. Dudin A. N., Shaban A. A., Klimenok V. I. Analysis of a $BMAP|G|1|N$ queue // International Journal of Simulation: Systems, Science and Technology. 2005. V. 6. № 1-2. P. 13–23.
4. Banik A. D., Ghosh S. Efficient computational analysis of non-exhaustive service vacation queues: $BMAP/R/1/N(1)$ under gated-limited discipline // Applied Mathematical Modelling. 2019. V. 68. P. 540–562.
5. Ghosh S., Banik A. D., Walraevens Joris, Bruneel Herwig. A Detailed Note on the Finite-Buffer Capacity Queueing System with Correlated Batch Arrivals and Batch-size-Phase-dependent Bulk-service // 4OR. 2021. doi.org/10.1007/s10288-021-00478-x
6. Kant L., Sanders W. H. Analysis of the distribution of consecutive cell losses in an ATM switch using stochastic activity networks // Computer Systems Science and Engineering. 1997. V. 12(2). P. 117–129.
7. Chydzinski A. On the distribution of consecutive losses in a finite capacity queue // WSEAS Trans. Circuits Syst. 2005. V. 4(3). P. 117–124.
8. De Boer P.T. Analysis and efficient simulation of queueing models of telecommunication systems. Ph.D. thesis, Centre for Telematics and Information Technology University of Twente. 2000.
9. Pacheco A., Ribeiro H. Consecutive customer loss probabilities in $M/G/1/n$ and $GI/M(m)/1/n$. In: Proceedings of Workshop on Tools for Solving Structured Markov Chains (SMCtools'06). Pisa, Italy. 2006.
10. Pacheco A., Ribeiro H. Consecutive customer losses in oscillating $GI^X/M/1/n$ systems with state dependent services rates // Annals of Operations Research. 2008. V. 162(1). P. 143-158.
11. Pacheco A., Ribeiro H. Consecutive customer losses in regular and oscillating $M^X/G/1/n$ systems // Queueing Systems. 2008. V. 58(2). P. 121-136.
12. Lee C.W., Andersland M.S. Consecutive cell loss controls for leaky-bucket admission systems. In: Global Telecommunications Conference, 1996. GLOBECOM'96.'Communications: The Key to Global Prosperity. IEEE 1996. V. 3. P. 1732-1738.
13. Вишневский В.М., Дудин А.Н., Клименок В.И. Стохастические системы с коррелированными потоками: Теория и применение в телекоммуникационных сетях. 2018. М.:Техносфера.
14. Neuts M. F. Matrix-Geometric Solutions in Stochastic Models. Baltimore:The Johns Hopkins University Press. 1981.

UDC: 621.396.49

Code Division Based on M-sequences and Its Optimization

D.S. Kukunin¹, A.A. Berezkin², R.V. Kirichek³

¹Bonch-Bruevich Saint-Petersburg State University of Telecommunications, 192232,
22 Bolshevikov Ave., build.1, St. Petersburg, Russia

²Bonch-Bruevich Saint-Petersburg State University of Telecommunications, 192232,
22 Bolshevikov Ave., build.1, St. Petersburg, Russia

³Bonch-Bruevich Saint-Petersburg State University of Telecommunications, 192232,
22 Bolshevikov Ave., build.1, St. Petersburg, Russia

coux@yandex.ru, aa.berezkin@mail.ru, kirichek@sut.ru

Abstract

The work contains a description of methods for creating orthogonal constructions based on recurrent sequences of maximum length. These structures are formed when the signal spectrum is spread by direct sequences and provide signal densification, as well as code division of channels. At the same time, an effective way of optimizing orthogonal structures is proposed, which should increase the energy efficiency of the signal.

Keywords: code division of channels, multiple access, M-sequence, dual basis, Galois field, field element

1. Introduction

Data transmission using tethered high-altitude unmanned telecommunication platforms ensures the construction of a new generation of networks, for which energy efficiency is of particular importance. Energy efficiency in wireless networks is achieved through various means, among which a special place is occupied by the technology of spread spectrum of signal, which in this work is used to organize the code division of channels.

The advantages of multiple access technology based on code division or code channel division over time and frequency division are obvious [1]. Suffice it to recall that TDMA (Time Division Multiple Access) and FDMA (Frequency Division Multiple Access) rely on physical quantities that, one way or another, are their common resource. At the same time, time is the most valuable resource, and the frequency range, as a rule, is regulated by legislative and legal restrictions.

Multiple access based on the CDMA (Code Division Multiple Access) is not a new technology, its foundations were laid in the last century. The CDMA civilian communication system, known as IS-95, used Walsh [2] codes as spreading the spectrum of sequences.

Walsh codes are orthogonal in nature. They are analogs of digital sinusoids built on the basis of Rademacher functions. The main disadvantage of Walsh sequences can be considered, first of all, an insufficiently "good" autocorrelation function, which does not allow them to fully withstand phase failures. This work suggests using structures based on recurrent M-sequences to expand the signal spectrum and further code compaction.

2. Forward and inverse M-sequences

The method of spread spectrum of the signal by a direct sequence, depending on the tasks being solved, uses various variants of the code combinations [3, 4, 5, 6, 7]. A good autocorrelation function reduces the probability of phase failures, at the same time a large code distance between combinations provides high noise immunity [8, 9].

The most important factor in the joint use of sequences suitable for spread spectrum is their orthogonality. It is this property that makes it possible to produce code compaction and transmit independent streams of information in a common frequency band.

By their nature, sequences of maximum length are quasi-orthogonal, but code constructions based on a linear sum of direct and inverse M-sequences are orthogonal with respect to them [10]. This allows you to combine them into a common frequency spectrum and transmit in one time interval.

The direct M-sequence $\{U\}$ can be constructed in various ways [11], for example, using a scheme based on a recurrent shift register with outputted adders. To construct an inverse M-sequence R , this polynomial is multiplied by $(x+1)$. Thus, the inverse sequence construction scheme for $P(x)=(x^4+x+1)(x+1)=x^5+x^4+x^2+1$ will have the form.

3. Orthogonal structures based on M-sequences

The linear sum of direct and inverse M-sequences is orthogonal to all M-sequences that are not included in this sum. For example, consider signal $S=(0 -2 +2 0 +2 0 -6 +6 +2 +2 -4 -4 0 +2 0)$, which includes four direct M-sequences with initial phases: $\varepsilon, \varepsilon^3, \varepsilon^8, \varepsilon^{12}$ and four inverses with phases: $\varepsilon^4, \varepsilon^5, \varepsilon^9, \varepsilon^{14}$.

Let's decompose all k -element sections of the signal S by the dual basis of the field $GF(2^k)$. The elements of the dual basis $\{\lambda\}$ are connected to the elements of the left power basis $\{\alpha\}$ via the trace function [12]:

$$T(\lambda_i \alpha_j) = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j. \end{cases} \quad (1)$$

Processing of the first section, as expected, forms a field element that corresponds to the sum of the phases of all M-sequences included in the signal S : $\varepsilon + \varepsilon^3 + \varepsilon^8 + \varepsilon^{12} + \varepsilon^4 + \varepsilon^5 + \varepsilon^9 + \varepsilon^{14} = \varepsilon^5$. One of the densest sections (+6 +2 +2 -4) in this case, it should form a field element shifted from the initial phase of ε^5 by 7 steps, that is, ε^{12} . Indeed, if the entire signal S is processed, it becomes clear that the energy levels of L include the same pairs of field elements (Table 1).

L	Number of the k -element section														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6				1	ε^4	ε^5	ε^{13}	ε^{14}							
5				1	ε^4	ε^5	ε^{13}	ε^{14}							
4				1	ε^4	ε^5	ε^{13}	ε^3	ε^4	ε^5	ε^{13}	ε^{14}			
3				1	ε^4	ε^5	ε^{13}	ε^3	ε^4	ε^5	ε^{13}	ε^{14}			
2	ε^5	ε^6	ε^7	ε^8	ε^9	ε^{10}	ε^{11}	ε^{11}	ε^{11}	ε^{12}	ε^6	ε^7	ε^2	ε^3	ε^4
1	ε^5	ε^6	ε^7	ε^8	ε^9	ε^{10}	ε^{11}	ε^{11}	ε^{11}	ε^{12}	ε^6	ε^7	ε^2	ε^3	ε^4
$\sum_{1,3,5}$	ε^5	ε^6	ε^7	ε^8	ε^9	ε^{10}	ε^{11}	ε^{12}	ε^{13}	ε^{14}	1	ε	ε^2	ε^3	ε^4
$\sum_{2,4,6}$	ε^5	ε^6	ε^7	ε^8	ε^9	ε^{10}	ε^{11}	ε^{12}	ε^{13}	ε^{14}	1	ε	ε^2	ε^3	ε^4

Table 1. The result of complete processing of the original signal by a dual basis

Thus, there is a variant of optimizing the compacted signal S in order to increase its energy efficiency. Removing even or odd layers is equivalent to halving the signal level. In fact, this is the averaging of energy between the direct and inverse M-sequences that are part of it. After this operation, the signal takes the form: $S_{\text{opt}} = (0 -1 +1 0 +1 0 -3 +3 +1 +1 -2 -2 0 +1 0)$.

Its processing at the reception by a dual basis will also make it possible to accurately distinguish the initial phase and all subsequent elements of the field. As will be shown later, determining the M-sequences included in the signal will not require bringing the S_{opt} to its original form, since it inherits all orthogonal properties from the original signal.

Consider an example with a longer polynomial $P(x) = x^8 + x^4 + x^3 + x^2 + 1$. Now the number of address combinations is 255 M-sequences. Let the direct combinations be used to transmit the symbol "0" and the inverse ones encode the symbol "1". At the same time, for convenience, we uniquely compare the channel numbers to

the degrees of the Galois field elements $\text{GF}(2^8)$, which form the initial phases of the corresponding address combinations. That is, channel 10 will use the M-sequence with the initial phase ε^{10} , channel 100 - ε^{100} and so on.

We organize the transmission of the characters "0" through randomly selected channels: 12, 46, 83, 95, 144, 168, 174, 196, 203, 241. Similarly, we use random channels to transmit characters "1": 3, 9, 38, 75, 102, 134, 156, 215, 226, 253. It can be represented graphically in comparison with the main signal (Fig. 1).

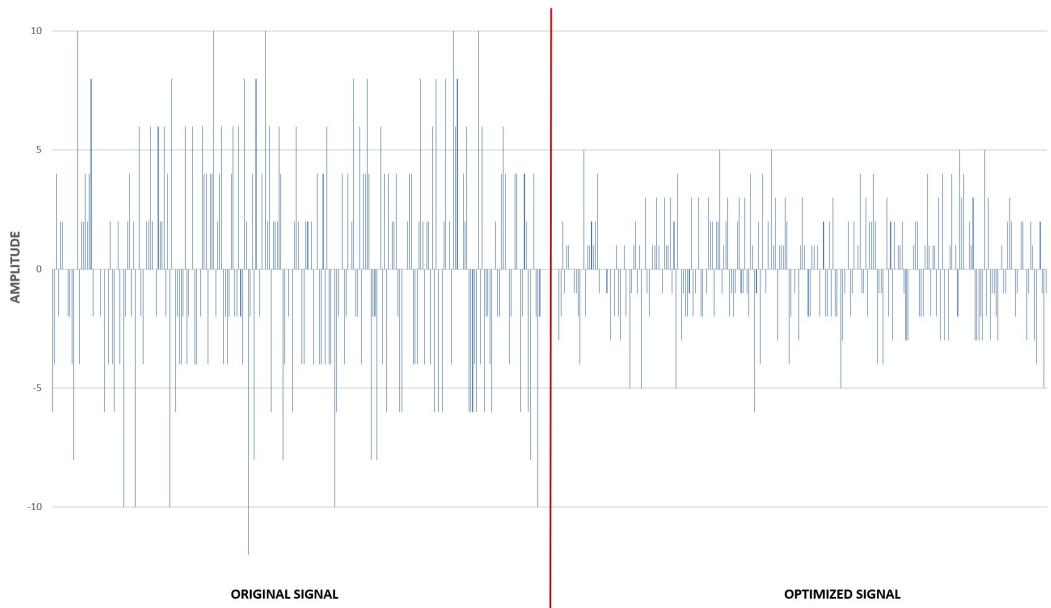


Fig. 1. A signal based on twenty M-sequences and its optimization

By sequentially processing k -element sections of S_{opt} with the dual basis of the Galois field $\text{GF}(2^8)$, you can make sure that a number of field elements appear, starting with the vector ε^{117} , which is the sum of the phases of twenty previously selected M-sequences. Thus, the equality is fulfilled:

$$\varepsilon^{117} = \varepsilon^{12} + \varepsilon^{46} + \varepsilon^{83} + \varepsilon^{95} + \varepsilon^{144} + \varepsilon^{168} + \varepsilon^{174} + \varepsilon^{196} + \varepsilon^{203} + \varepsilon^{241} + \varepsilon^3 + \varepsilon^9 + \varepsilon^{38} + \varepsilon^{75} + \varepsilon^{102} + \varepsilon^{134} + \varepsilon^{156} + \varepsilon^{215} + \varepsilon^{226} + \varepsilon^{253}.$$

It is obvious that interference affecting the signal should lead to distortion of the results of its processing at reception. But, considering that S_{opt} is processed by the dual basis as equidistant code, it can be argued about its high noise immunity.

At reception, the optimized signal S_{opt} is multiplied by all address sequences in the same way as the original signal S . Previously, we associated channel numbers

with the degree of field elements that generate address sequences. We will allocate information, for example, for 95, 156 and 62 channels. Thus, when multiplying S_{opt} by M-sequences U_{95} , U_{156} , U_{62} with phases ε^{95} , ε^{156} and ε^{62} , we should get the normalized values +1 (the direct address sequence was transmitted), -1 (the inverse address sequence was transmitted) and 0 (nothing was transmitted), accordingly. Taking into account the features of the orthogonal structure we have constructed and the fact that the signal amplitude has been halved, we will normalize the result of scalar multiplication by the value $(n+1)/2=128$, where n is the period of M-sequences.

As expected, reducing the amplitude of the original signal by half did not affect the result of detecting information in the channels. The direct M-sequence was indeed transmitted via channel 95, the inverse M-sequence was transmitted via channel 156, no information was transmitted via channel 62, therefore, as a result of multiplication by the address sequence U_{62} , zero was obtained.

4. Conclusion

Code constructions based on linear sums of recurrent sequences of maximum length are multilayer orthogonal structures that are processed quite efficiently by the dual basis of the Galois field. The result of the sequential decomposition of their sections on a dual basis are double energy layers that contain the same elements of the field. Thus, the sum of the field vectors of even and odd levels is the same and equal to the field element, which, in turn, is the sum of the phases of the M-sequences that make up the multilayer structure.

The properties of the maximum length sequences considered in this paper are capable of providing constant control over the transmission of a compacted signal. A phase failure or a jump in the signal level will lead to a break in a number of field elements obtained as a result of processing the code structure with a dual basis. At the same time, orthogonality at the reception will allow you to unambiguously divide the signal into components.

The paper also suggests the idea of increasing energy efficiency by reducing the signal level by half. This procedure is designed to eliminate code redundancy. At the same time, it will not affect the result of determining the phase of the signal during reception.

5. Acknowledgments

The study was financially supported by the Russian Science Foundation within of scientific project No. 22-49-02023 "Development and study of methods for obtaining the reliability of tethered high-altitude unmanned telecommunication platforms of a new generation"

REFERENCES

1. Viterbi A. CDMA: Principles of Spread Spectrum Communication. - Englewood Cliffs, NJ.: Prentice Hall, 1995.
2. Tanenbaum Andrew S. Computer networks / Andrew S. Tanenbaum, David J. Wetherall. - 5th ed. Pearson Education, Inc.: Prentice Hall, 2011.
3. Don Torrieri Principles of Spread-Spectrum Communication Systems - 4th ed. Springer Nature, 2018. – 727 p. <https://doi.org/10.1007/978-3-319-70569-9>
4. D. A. Visan Direct Sequence Spread Spectrum Communication Module for Efficient Wireless Sensor Networks / D. A. Visan, M. Jurian, I. Lita, L. M. Ionescu, A. G. Mazare // Electronics Computers and Artificial Intelligence (ECAI) 2019 11th International Conference on, pp. 1-4, 2019.
5. Aya Y. Khudhair Reduction of the Noise Effect to Detect the DSSS Signal using the Artificial Neural Network / Aya Y. Khudhair, Rajaa A. Abd Khalid // Information Technology and Science (BICITS) 2021 1st Babylon International Conference on, pp. 185-188, 2021.
6. Yuyao Shen Analysis of the Code Phase Migration and Doppler Frequency Migration Effects in the Coherent Integration of Direct-Sequence Spread-Spectrum Signals / Yuyao Shen, Ying Xu // Access IEEE, vol. 7, pp. 26581-26594, 2019.
7. Qiu Z. A Blind Despreadening and Demodulation Method for QPSK-DSSS Signal With Unknown Carrier Offset Based on Matrix Subspace Analysis / Z. Qiu, H. Peng, T. Li // IEEE Access. – 2019. – Vol. 7. – P. 125700-125710.
8. Patent RU2621181C1. Russian Federation. H03M 13/15, H04L 7/02. Cycle Synchronization Method with Dynamic Addressing Recipient / Kognovitskij O.S., Vladimirov S.S., Kukunin D.S., Lapshov D.Y. – № 2016121944. Registration date: 02.06.2016; date of publication: 31.05.2017.
9. Kukunin D. Phasing in Asynchronous Data Transmission System using M-sequences / D. Kukunin, A. Berezkin, R. Kirichek, I. Pantaleimonov // ICFNDS 2021: The 5th International Conference on Future Networks & Distributed Systems. – December 2021. – P. 516–521. <https://doi.org/10.1145/3508072.3508178>
10. Kukunin D.S., Berezkin A.A., Kirichek R.V. The use of Phantom Channels as Catalysts for Enhancing the Orthogonal Properties of M-sequences in CDMA // Trudy Nauchno-Issledovatel'skogo Instituta Radio. 2022. № 1. P. 37-47. <https://doi.org/10.34832/NIIR.2022.8.1.004>.
11. Kognovickij O.S. *Dvojstvennyj Bazis i ego Primenenie v Telekommunikacijah*. – SPb.: Link, 2009. – 424 p.
12. Kukunin D. Dependence of the Structural Properties of the Dual Basis on the type of the Characteristic Polynomial // Informacionnye Tehnologii i Telekomunikacii. – 2019. – V. 7. № 1. – P. 41-51. <https://doi.org/10.31854/2307-1303-2019-7-1-41-51>.

UDC: 004.032.26

Efficient data coding methods based on neural networks

A.A. Berezkin¹, D.S. Kukunin², A.V. Slepnev³, R.V. Kirichek⁴

¹St. Petersburg State University of Telecommunications named after Prof. M.A. Bonch-Bruevich, 22 Bolshevikov Ave. 1, St. Petersburg, Russia

²St. Petersburg State University of Telecommunications named after Prof. M.A. Bonch-Bruevich, 22 Bolshevikov Ave. 1, St. Petersburg, Russia

³St. Petersburg State University of Telecommunications named after Prof. M.A. Bonch-Bruevich, 22 Bolshevikov Ave. 1, St. Petersburg, Russia

⁴St. Petersburg State University of Telecommunications named after Prof. M.A. Bonch-Bruevich, 22 Bolshevikov Ave. 1, St. Petersburg, Russia

aa.berezkin@mail.ru, coux@yandex.ru, avs99.98@gmail.com, kirichek@sut.ru

Abstract

This article discusses a neural network-based compression algorithm using noise-correcting codes. The use of this algorithm has a number of advantages, on the one hand, the noise-resistant code allows to get rid of potentially high overheads provoked by the use of a neural network, lowering the required value of model accuracy to the value determined by the correctability of the used code. On the other hand, a trained neural network allows data compression without prior transformations. .

Keywords: *Autoencoder, data compression, neural networks*

1. Introduction

Nowadays, the emergence of new types of communication networks, particularly, networks using tethered and autonomous UAVs, has led to the development of different data transmission methods [1, 7, 8]. As expected in the future networks, it is required to consider having an ultra-low latency and high bandwidth for some applications such as augmented/virtual reality or tactile internet [6]. The modern stage of telecommunications development imposes stringent requirements for speed, delay and reliability in transmitting and processing information.

One of the approaches to increase the speed of data transmission, and hence to eliminate delays, is the compression of information in communication channels on the fly. This approach requires the development of new information processing

architectures that allow data processing in parallel. Neural networks by virtue of their structure can fulfill the set requirements.

Neural networks are actively used in image compression tasks [2], since the generalization ability of this method allows to achieve lossy compression. Theoretically, it is possible to achieve full data recovery [3], but in practice developers face high overhead costs associated with an increase in the occupied space of the network itself, as well as, with an increase in the computational complexity of the algorithm. The proposed codec structure is shown in figure 1, where the neural network part of the codec is highlighted by a rectangle.

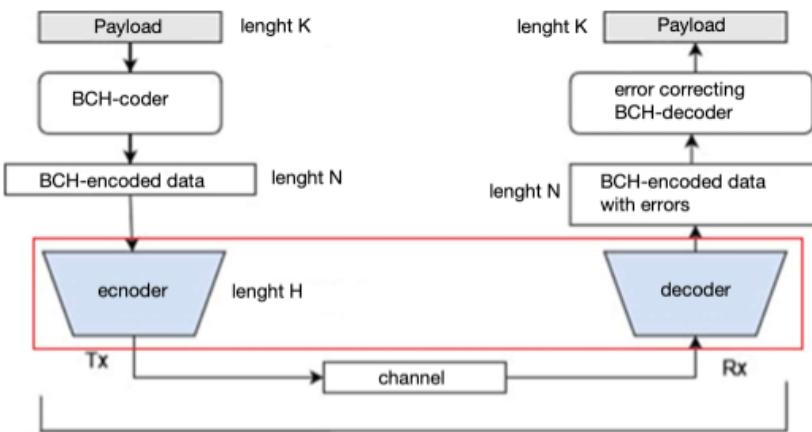


Fig. 1. Neural network codec architecture

This architecture consists of several independent processing units:

- Interference code encoder and decoder;
- A compression algorithm based on a neural network;

The use of this architecture has several advantages, on the one hand, the error-correcting code [3, 4] allows to get rid of the potentially high overheads induced by the use of the neural network, lowering the required value of the final model accuracy to the value determined by the corrective capability of the used noise-correcting code. On the other hand, a trained autoencoder neural network allows data compression without prior transformations.

In this paper we investigated the operation of the autoencoder using several error-correcting codes of different multiplicity. The following error-correcting codes were used: (31, 21), (31, 16) and (31, 5).

These codes have different correction ability, as well as different number of information bits, which will allow to better explore the capabilities of the neural

network and determine the dependence of the hidden state size not only on the number of guaranteed correctable errors, but also to determine the influence of the code distance on the generalization ability of the autoencoder neural network.

The criterion for comparing different architectures is the fact of lossless data recovery from the hidden state. This metric can be described as the accuracy of the neural network and defined as the fraction of correctly recovered symbols relative to the whole set.

For lossless data recovery, the accuracy should be equal to one. The accuracy shows the quality of the neural network with respect to the whole data set, not paying attention to errors in specific code combinations. In practice, there may be situations in which the neural network will perform well relative to the full data set, but some of the code combinations cannot be recovered in their original form. To control the accuracy of the autoencoder at the level of a particular code combination, an additional metric was introduced, defined as the minimum accuracy of the neural network relative to the code combination.

The following formula was used to calculate the lower bound of autoencoder accuracy:

$$CodeWiseAcc_{min} = \frac{n - t}{n}, \quad (1)$$

where $CodeWiseAcc_{min}$ is the lower bound of the required accuracy, n is the number of elements of the code sequence containing check and information symbols, t is the number of guaranteed error correction by the error code.

The boundaries of acceptable accuracy of the autoencoder when using anti-interference codes to provide lossless compression are presented in the table 1.

Interference-resistant code	Lower limit of tolerable accuracy of autoencoder
BCH (31, 21)	0,9355
BCH (31, 16)	0,9032
BCH (31, 5)	0,7580

Table 1. Limits of acceptable neural network accuracy

2. Experimental results

2.1. BCH-based codec (31, 5). During the study it was found that the use of a noise-correcting code made it possible to represent the initial combination in the form of a vector of length 2, while the model without the mechanism used showed itself worse, achieving lossless compression with a hidden state length equal to 3. The results are given in the table 2.

Model reconstruction accuracy	The size of the hidden state vector		
Using anti-jamming code	1	2	3
False	0,78125	0,8999999976	1
True	0,637499988	1	1

Table 2. results based on the BCH codec (31, 5)

Thus, in the course of simulation with the use of interference-free coding the compression coefficient was achieved equal to 0.4. On the other hand, without the use of noise-correcting code compression factor is equal to 0.6. Thus, increase of code redundancy by 83

Despite the fact that the neural network autoencoder using the noise-free coding algorithm achieved a lossless compression effect, the error function values differ slightly, which confirms the theory of achieving full recovery of the original data by error correction.

2.2. BCH-based codec (31, 16). During the study it was found that the use of a noise-correcting code made it possible to represent the initial combination in the form of a vector of length 7, while the model without the mechanism used showed itself worse, achieving lossless compression with a hidden state length equal to 8. The results are given in table 3.

Model reconstruction accuracy	The size of the hidden state vector		
Using anti-jamming code	1	2	3
False	0,793	0,778	1
True	0,886	1	1

Table 3. results based on the BCH codec (31, 16)

Thus, during the simulation with the use of interference-free coding the compression ratio was achieved equal to 0.43. On the other hand, without the use of noise-correcting code compression factor is equal to 0.5. Thus, a 52% increase in code redundancy entails a 12.5% decrease in the amount of data in the link.

Starting from 37567 epoch the neural network autoencoder constructed with the use of the interference-resistant code reaches the accuracy equal to one, which indicates lossless data compression. Neural network autoencoder without use of interference-free coding achieves this value at step 8863, which is much smaller. This fact is explained by the fact that the neural network autoencoder without the use of noise-correcting code recovers only 16 information symbols against 31.

Thus, the proposed implementation of the neural network codec allows to get rid of less redundancy in the channel while using the code with less correcting power. The reason of this fact may be the polynomial growth of the training set with the increase of information symbols, which requires to change the number of training parameters of the neural network in the higher side.

2.3. Code based on BCH (31, 21). During the study of the neural network autoencoder using the interference-resistant BSH code (31, 21), a contradiction was identified. A neural network autoencoder without the use of a noise-correcting code, having a hidden vector size equal to the length of the original sequence could not achieve the necessary accuracy for lossless compression. Table 4 shows the accuracy values depending on the hidden layer length and the condition of using the interference-resistant code.

Model reconstruction accuracy	The size of the hidden state vector						
Using anti-jamming code	7	9	11	12	13	15	21
False	0,67	0,74	0,71	-	0,77	-	0,72
True	0,39	0,52	0,52	0,68	-	0,716	0,77

Table 4. The results obtained based on the BCH codec (31, 21)

Thus we can conclude that the structure of the neural network autoencoder, does not have sufficient complexity to provide lossless data compression for the given number of information symbols and the size of the training set. Therefore to study this code and codes with larger length of information bits it is necessary to use more complex architectures with a larger number of trainable parameters.

3. Conclusion

The following results were obtained during the analysis: using the FFT code (31, 5) it was possible to achieve a compression ratio equal to 0.4, which will reduce the data volume in the communication channel by 33% when modifying the architecture based only on the neural network autoencoder. When using the BCH code (31, 16), a compression ratio of 0.43 was achieved, which will reduce the amount of data in the communication channel by 12.5% when modifying the architecture based only on the neural network autoencoder. Also, the study revealed disadvantages that must be taken into account in the system design. During the comparison of the obtained accuracy when using codes with different minimum code distances, it was determined that the data structure formed by cyclic codes allows the neural network to better determine the weighting coefficients and restore the input sequence. The results of this comparison showed that a larger minimum code distance of the input sequence

leads to greater relative accuracy of the model. This means that when the minimum code distance increases, the recovery potential of the neural network autoencoder increases, which can be used to solve problems in related areas.

4. Acknowledgement

The study was financially supported by the Russian Science Foundation within of scientific project No. 22-49-02023 "Development and study of methods for obtaining the reliability of tethered high-altitude unmanned telecommunication platforms of a new generation"

REFERENCES

1. A. Koucheryavy, A. Vladyko, R. Kirichek, "State of the art and research challenges for public flying ubiquitous sensor networks", Internet of Things, Smart Spaces, and Next Generation Networks and Systems, 2015, pp. 299-308.
2. Wu Y. et al. Deep Image Compression with Latent Optimization and Piece-wise Quantization Approximation //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2021. – P. 1926-1930.
3. Berezkin, A., Zadorozhnyaya, A., Kukunin, D., Matveev, D., Kraeva, E., Models and Methods for Decoding of Error-Correcting Codes based on a Neural Network// International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, 2021-October, P. 230 – 235.
4. S. Ahmed, Linear block code decoder using neural network // IEEE International Joint Conference on Neural Networks (IJCNN'08), 2008, P. 1111–1114.
5. A. Claytus Vaz, G. Nayak, D. Nayak, Hamming Code Performance Evaluation using Artificial Neural Network Decoder // 15th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, Romania, June 2019, P. 37–40.
6. M. Al-gaashani, M. S. A Muthanna, K. Abdulkadir, A. Muthanna, and R. Kirichek, Intelligent System Architecture for Smart City and its Applications Based Edge Computing // 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2020, P. 269-274.
7. S. Vladimirov, R. Kirichek, and V. Vishnevsky, Network Coding for the Interaction of Unmanned Flying Platforms in Data Acquisition Networks //The 4th International Conference on Future Networks and Distributed Systems (ICFNDS), 2020, P. 1-7.
8. S. Vladimirov, V. Vishnevsky, A. Larionov, R. Kirichek, The Model of WBAN Data Acquisition Network Based on UFP // International Conference on Distributed Computer and Communication Networks (DCCN-2020), 2020, P. 220-231.

UDC: 303.717,004.421

The estimation of microchip testing process duration based on extended fault injection method

O.M.Brekhov¹ and A.V.Klimenko²

¹Moscow Aviation Institute, Volokolamskoye shosse 4, Moscow, Russia

²Moscow Aviation Institute, Volokolamskoye shosse 4, Moscow, Russia

obrekhov@mail.ru, a.v.klimenko@mai.ru

Abstract

The need of adding some kind of fault tolerance to modern microchip design is of great importance today, both for aerospace and sea level applications. It is mostly explained by the increased sensitivity to negative environmental factors due to technology shrinks and because of the need of chip yield increasing in the presence of manufacturing defects. This article discusses an algorithm for assessing microchip design fault tolerance, based on the extended fault injection method.

1. Introduction

The need of adding some kind of fault tolerance to modern microchip design is of great importance today, both for aerospace and sea level applications. It is mostly explained by the increased sensitivity to negative environmental factors due to technology shrinks and because of the need of chip yield increasing in the presence of manufacturing defects. This fact determines the need to develop new methodologies of microchip testing that also implement fault tolerance assessment.

2. Previous work

In [1, 2] the chip modeling methodology using extended fault injection method is proposed. The methodology allows to evaluate the fault tolerance of the microchip design for the given parameters of external influences. The methodology implementation assumes FPGA based hardware-software complex (HSC) usage. The HSC consists of workstation connected to several extension boards via PCIe (see figure 1).

Such hardware-software organization is very common in FPGA-based fault injection methodologies: the workstation contains software complex that implements user interface, test results processing, input data generation and response collection for

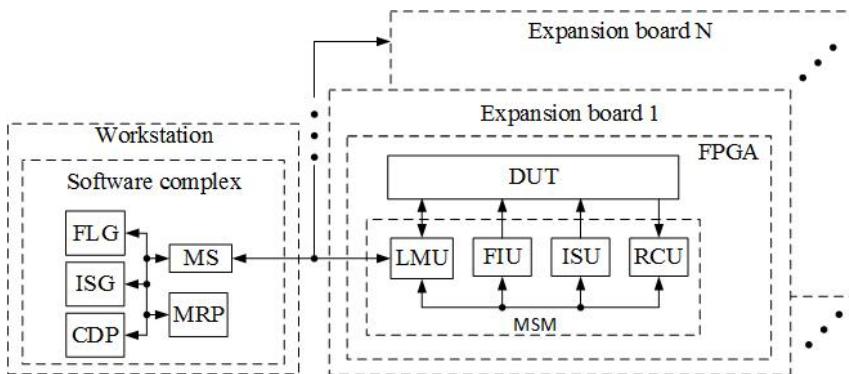


Fig. 1. The HSC functional block diagram

the device under test (DUT). The DUT is implemented in FPGA of the extension board. The software complex consists of: fault list generation component (FLG), input stimuli generation component (ISG), chip design processing component (CDP), modeling results processing component (MRP), modeling support component (MS). The Distinctive feature of the extended fault injection method is so-called modelling support microkernel (MSM) usage. The MSM communicates directly with DUT inputs and outputs and consists of these hardware units: local management unit (LMU), fault injection unit (FIU), input stimuli supply unit (ISU), response collection unit (RCU). The FLG accordingly to the extended fault injection method contains three software models, which allow modeling of the processes that could cause a fault during chip operation period. The concept of these three models usage is an another feature of the extended fault injection method. These models are: External Influences Model (EIM), Threats Occurrence Model (TOM) and Fault Localization Model (FLM). The EIM model sets parameters of negative environmental impacts, that could damage microchip during its operation period. Space radiation is an example of such negative environmental impact. It is known as harmful environmental factor not only for aerospace, but also for sea-level electronics [1]. The EIM model uses charged particles fluxes parameters as input data when space radiation is considered as primary harmful environmental factor. The EIM model output contains charged particles fluxes parameters that directly affect the microchip (taking into account the applied radiation protection). The TOM model uses the EIM output as input data, jointly with microchip topology. The TOM model output contains parameters of physical characteristics changes in the microcircuit that could cause malfunctioning. The FLM model generates fault list based on the TOM output. Faults from the fault list would be injected in the DUT during fault tolerance testing process. Thus, FLM converts data on changes in the internal environment of the chip into fault types.

that could be hardware injected to the DUT and localization of places where they should be injected during testing process. To provide feasibility of injecting faults into the DUT (implemented in FPGA), this project must first be modified. The modification consists in replacing the elements of the DUT with functionally identical ones (saboteur modules), but containing failure models that can be activated during the testing process (see Figure 2). Figure 2 schematically shows the elements of the microchip project. Replaceable elements are marked with a dotted line. As can be seen from the figure, the DUT after modification has KN additional inputs, called "fault activation inputs". Here k is the number of fault activation inputs for one saboteur module, N is the number of DUT elements to be replaced. The k value depends on the coding method used for the failure activation signals. Thus we could say that aforementioned hardware and software parts of the HSC form hardware-software environment for microchip project modification (HSEPM) which adds the ability to inject faults into it.

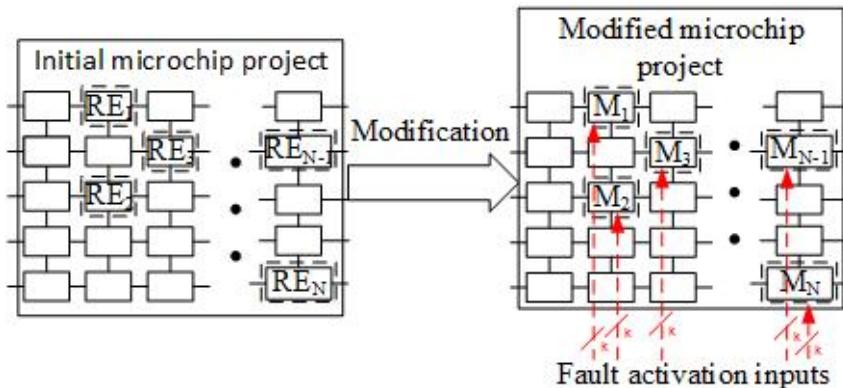


Fig. 2. Microchip project modification process. RE_i -replaceable module; SM_i -saboteur module, corresponding to RE_i .

3. The fault tolerance assessment algorithm

Due to the fact that HSEPM implies modification of the original microchip design, the testing algorithm must ensure the functional correspondence of the original and modified designs. Moreover, the MSM usage in the extended fault injection method imposes additional requirements on the used FPGA capacity. It could lead to limitations when using this method for testing large-scale microchip designs with a large number of elements. This article discusses an algorithm for assessing microcircuit design fault tolerance, based on the extended fault injection method. The algorithm could be applied for testing process of the large-scale microchip designs Let's consider

- 10) $j = 0$.
- 11) $j = j + 1$
- 12) Testing functional compliance of the initial project and modified project using “detalization variant” with index j . Lets assume $t_M P_{ij}$. – is the time period needed for his process to complete.
- 13) If $j < m$ than jump to 11. Jump to 9 otherwise.

Thus, the total time required for the fault tolerance evaluation process is determined by the formula:

$$\sum_{i=1}^n (t_I P_i + t_M P_i + \sum_{l=1}^q (t_M P_i^l + t_K^l)) + \sum_{j=1}^m (t_M P_{ij})$$

, where q is number of needed HSEMP corrections.

4. Conclusion

The algorithm for microchip design fault tolerance assessment is proposed. It is based on an extended fault injection method. The algorithm assumes the possibility of selecting functional modules in the microchip design, each of which performs a specific function. The fault tolerance evaluation process within the proposed approach is performed separately for each module, and sequentially for all modules. This approach determines the potential applicability of the proposed algorithm for large chip designs testing process, as it allows to bypass the capacity limitations of the FPGAs used within the extended fault injection method.

REFERENCES

1. O. Brekhov, A. Klimenko Hardware-software simulation complex for FPGA-prototyping of fault-tolerant computing systems // Communications in Computer and Information Science V. 678. P. 72–86.
2. O. Brekhov, K. Kordover, A. Klimenko and M. Ratnikov FPGA prototyping with advanced fault injection methodology for tolerant computing systems simulation // Distributed Computer and Communication Networks V. 601. P. 208–223.
3. J. Barak, N. M. Yitzhak SEU Rate in Avionics: From Sea Level to High Altitudes // IEEE TRANSACTIONS ON NUCLEAR SCIENCE, V. 62. P. 3369–3380.

Научное издание

**РАСПРЕДЕЛЕННЫЕ КОМПЬЮТЕРНЫЕ
И ТЕЛЕКОММУНИКАЦИОННЫЕ СЕТИ:
УПРАВЛЕНИЕ, ВЫЧИСЛЕНИЕ, СВЯЗЬ
(DCCN-2022)**

Издание подготовлено в авторской редакции

Технический редактор *Н.А. Ясько*

Компьютерная верстка *Д.В. Козырев*

Дизайн обложки *Д.В. Козырев, М.В. Рогова*

Подписано в печать 22.09.2022 г. Формат 70×100/16. Печать офсетная.
Усл. печ. л. 35,48. Тираж 100 экз. Заказ 1147.

Российский университет дружбы народов
115419, ГСП-1, г. Москва, ул. Орджоникидзе, д. 3

Типография РУДН
115419, ГСП-1, г. Москва, ул. Орджоникидзе, д. 3.
Тел.: 8 (495) 955-08-74. E-mail: publishing@rudn.ru