Performance Evaluation of Computer Systems

Course overview

Tran, Van Hoai (hoai@hcmut.edu.vn)

Faculty of Computer Science & Engineering
HCMC University of Technology

2023-2024/Semester 1

Outline

1. Motivating examples

2. Basic concepts
   - Performance of a system
   - How to know system performance?

3. Course overview

4. Case-study
   - Supply chain by queueing network
   - Shortest path algorithms
   - “Art” in performance evaluation

Outline

1 Motivating examples

2 Basic concepts
- Performance of a system
- How to know system performance?

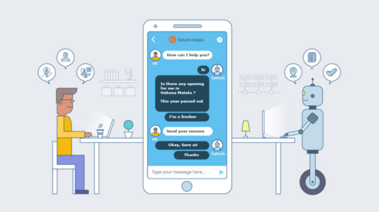
3 Course overview

4 Case-study
- Supply chain by queueing network
- Shortest path algorithms
- “Art” in performance evaluation

AI example

Chatbot

How to measure the quality of a commercial chatbot ?

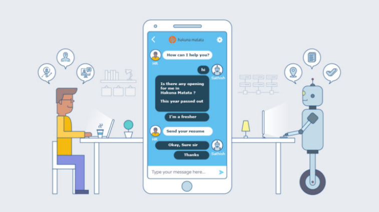


(source: Internet)

AI example

Chatbot

How to measure the quality of a commercial chatbot ?



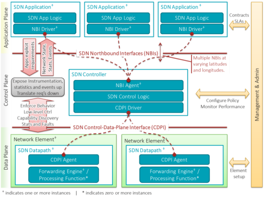
(source: Internet)

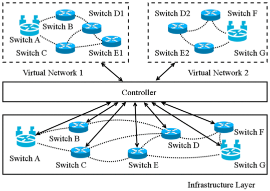
- **Self-service rate**: percentage of user sessions that did not end with a contact action after using the bot.
- **Performance rate**: number of correct answers divided by the number of active sessions (a correct answer is an answer suggested by the bot and clicked by the user in case of multiple choices – or opened instantly in case of strong semantic matching).
- **Usage rate per login**: volume of active user sessions on the chatbot. To balance out with the average number of sessions on your website.
- **Bounce rate**: volume of sessions where the chatbot was opened but not used
- **Satisfaction rate**: average grade given when evaluating the chatbot's answers (to balance out with the evaluation rate).
- **Evaluation rate**: percentage of user sessions that have given an evaluation of the chatbot's answers at least once.
- **Average chat time**: allows you to evaluate your users' interest for your chatbot.
- **Average number of interactions**: used to evaluate the Customer Effort Score on the chatbot and must be correlated to the satisfaction rate. If the latter is very low, the bot may be engaging the users in too many branches and steps to meet their needs. In this case, a resolution can be to correct the decision trees or knowledge base architecture.
- **Goal completion rate**: in case your bot contains targeted actions like CTAs, a form or some cross-selling, that is the rate of users who have reached that specific action through the chatbot.
- **Non-response rate**: the amount of times the chatbot has failed to push some content following a user question (due to lack of content or misunderstanding).

Software-Defined Network (SDN)

What is SDN? (Wiki)

SDN architectures decouple network control and forwarding functions, enabling the network control to become directly programmable and the underlying infrastructure to be abstracted from applications and network services.

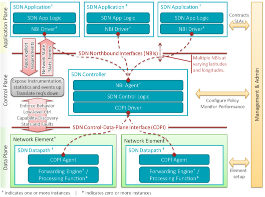

Application Plane
Control Plane
Data Plane
SDN Application$^+$
SDN App Logic
NBI Driver$^+$
SDN Application$^+$
SDN App Logic
NBI Driver$^+$
SDN Application$^+$
SDN App Logic
NBI Driver$^+$
Contracts SLAs
Apps explicit requirements
Network State, Stats, Events
SDN Northbound Interfaces (NBIs)
Multiple NBIs at varying latitudes and longitudes.
Expose Instrumentation, statistics and events up Translate req's down
SDN Controller
NBI Agent$^+$
SDN Control Logic
CDPI Driver
Configure Policy Monitor Performance
Enforce Behavior Low-level Ctrl Capability Discovery Stats and Faults
SDN Control-Data-Plane Interface (CDPI)
Network Element$^+$
SDN Datapath$^+$
CDPI Agent
Forwarding Engine$^+$ / Processing Function*
Network Element$^+$
SDN Datapath$^+$
CDPI Agent
Forwarding Engine$^+$ / Processing Function*
Element setup
Management & Admin
$^+$ indicates one or more instances | * indicates zero or more instances



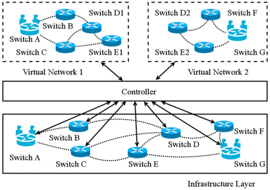
Software-Defined Network (SDN)

What is SDN? (Wiki)

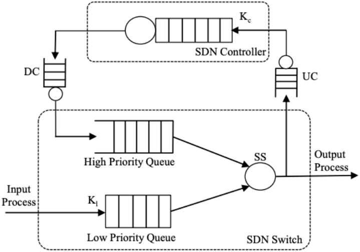
SDN architectures decouple network control and forwarding functions, enabling the network control to become directly programmable and the underlying infrastructure to be abstracted from applications and network services.



⁺ indicates one or more instances | * indicates zero or more instances



Performance of SDNs can be modeled as PQ-based architecture.



Can we train an adaptive controller faster?

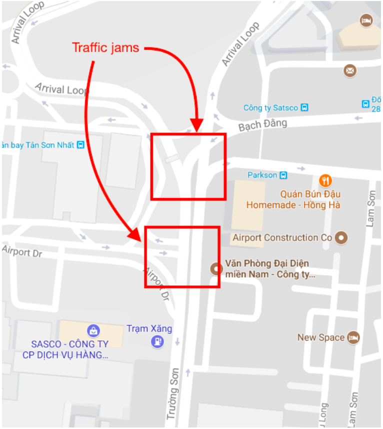
AlphaGo Zero

By playing games against itself, AlphaGo Zero surpassed the strength of AlphaGo Lee in three days by winning 100 games to 0, reached the level of AlphaGo Master in 21 days, and exceeded all the old versions in 40 days.

- See a tracking video of isolated intersection in Vietnam
- We can train a reinforcement learning traffic light controller from scratch, and better than state-of-the-art controllers.

Traffic in Tan Son Nhat Airport nearby

Before 8/2/2017



source: Google maps

source: cafebiz, vtv

Traffic in Tan Son Nhat Airport nearby

After 3/7/2017

- Investment: 242 billions VND
- Construction time: 8/2/2017 - 3/7/2017

Traffic in Tan Son Nhat Airport nearby

After 3/7/2017

- Investment: 242 billions VND
- Construction time: 8/2/2017 - 3/7/2017

source: vietnamnet.vn

Traffic in Tan Son Nhat Airport nearby

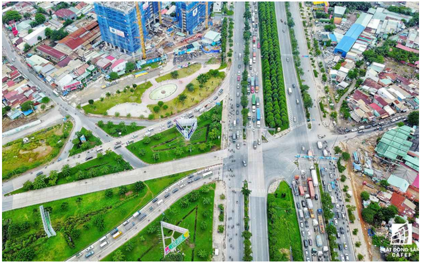
After 3/7/2017

- Investment: 242 billions VND
- Construction time: 8/2/2017 - 3/7/2017

source: vietnamnet.vn

source: vnexpress.vn, thanhnien.vn

Traffic in An Phu intersection



ĐƯỜNG MAI CHÍ THỌ
CAO TỐC LONG THÀNH - DẦU GIÂY
ĐƯỜNG LƯƠNG ĐỊNH CỦA

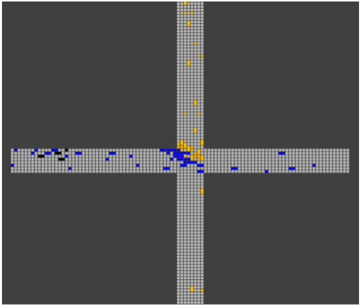
Can we know which one is better
in quantitative manner
right now?

What happened with the chosen solution ?

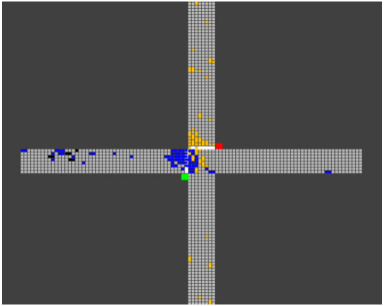
- We are here not transportation experts
  ⇒ do not discuss about which transportation methods/approaches are reasonable.
- There are something wrong in the process to choose builidng the overpass

Mixed traffic in Vietnam

No signalised traffic control
⇒ JAM, JAM, JAM



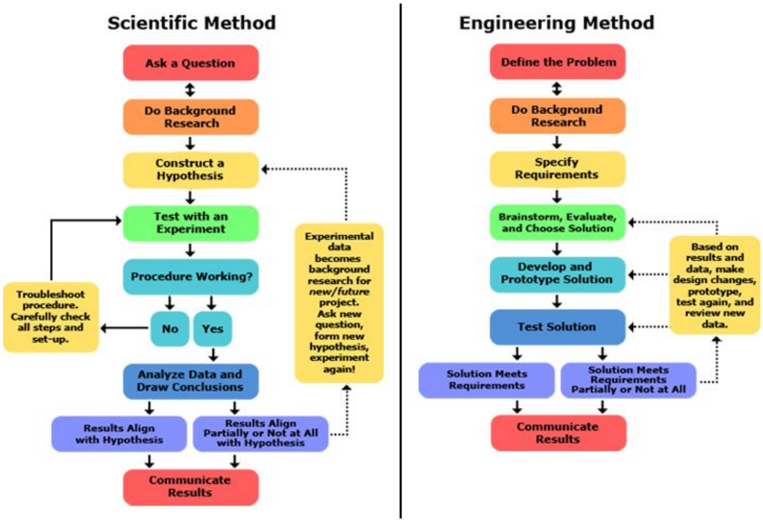
Signalised traffic control
⇒ ...and still JAM, JAM, JAM



Group discussion

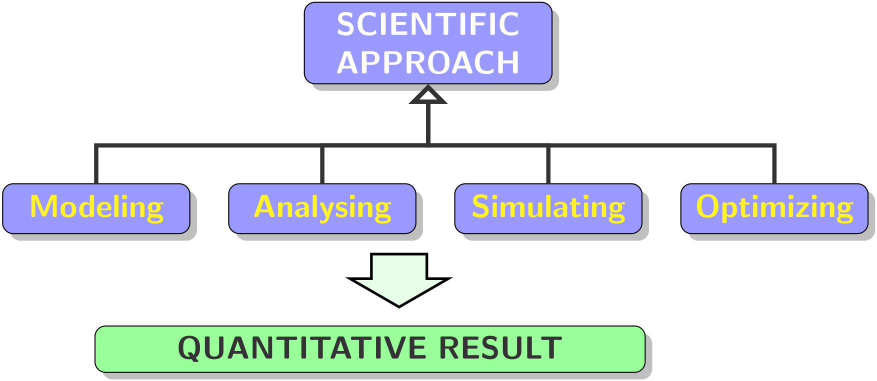
List possible main reasons of the jams at an intersection in the Vietnam big cities.

What were not performed to obtain the solution ?



Scientific Method
Ask a Question
Do Background Research
Construct a Hypothesis
Test with an Experiment
Procedure Working?
No
Yes
Troubleshoot procedure. Carefully check all steps and set-up.
Analyze Data and Draw Conclusions
Results Align with Hypothesis
Results Align Partially or Not at All with Hypothesis
Communicate Results
Experimental data becomes background research for *new/future* project. Ask new question, form new hypothesis, experiment again!
Engineering Method
Define the Problem
Do Background Research
Specify Requirements
Brainstorm, Evaluate, and Choose Solution
Develop and Prototype Solution
Test Solution
Solution Meets Requirements
Solution Meets Requirements Partially or Not at All
Communicate Results
Based on results and data, make design changes, prototype, test again, and review new data.

Quantitative approaches



SCIENTIFIC APPROACH

Modeling

Analysing

Simulating

Optimizing

QUANTITATIVE RESULT

Outline

1. Motivating examples

2. Basic concepts
   - Performance of a system
   - How to know system performance?

3. Course overview

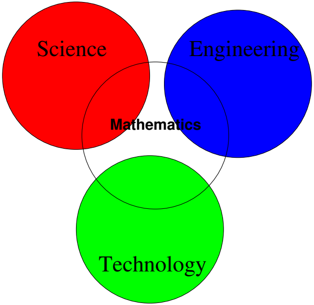
4. Case-study
   - Supply chain by queueing network
   - Shortest path algorithms
   - “Art” in performance evaluation

Engineering and Mathematics

What is an engineer ?

An engineer is a professional practitioner of engineering, concerned with applying scientific knowledge, mathematics and ingenuity to develop solutions for technical problems.

(wikipedia.org)

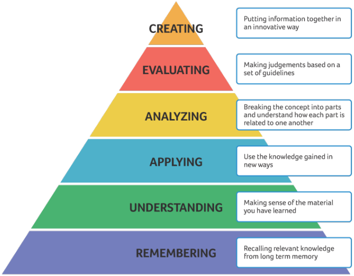


What is performance evaluation ? (1)

Performance evaluation aims at predicting a system's behavior in a quantitative manner.

What is performance evaluation ? (1)

Performance evaluation aims at predicting a system's behavior in a quantitative manner.



CREATING — Putting information together in an innovative way

EVALUATING — Making judgements based on a set of guidelines

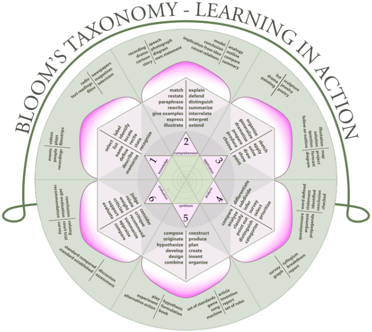
ANALYZING — Breaking the concept into parts and understand how each part is related to one another

APPLYING — Use the knowledge gained in new ways

UNDERSTANDING — Making sense of the material you have learned

REMEMBERING — Recalling relevant knowledge from long term memory

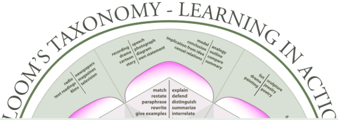
source: blog.newsela.com


BLOOM'S TAXONOMY - LEARNING IN ACTION
1 knowledge
2 comprehension
3 application
4 analysis
5 synthesis
6 evaluation
match restate paraphrase rewrite give examples express illustrate
explain defend distinguish summarize interrelate interpret extend
compose originate hypothesize develop design combine
construct produce plan create invent organize
label list name define describe memorize select identify locate recite state recognize
organize generalize dramatize prepare produce choose apply solve draw show paint sketch
judge relate weight criticize support evaluate consider critique recommend summarize appraise compare
differentiate subdivide compare analyze classify point out distinguish categorize infer survey select prioritize

source: wikipedia.org

What is performance evaluation ? (1)

Performance evaluation aims at predicting a system's behavior in a quantitative manner.



Examples

- The number of stations that can be connected to a LAN and still maintain a reasonable average frame delay and throughput ?
- The fraction of calls that are blocked on outgoing lines of a company's telephone system and how much improvement we can get if an extra line is added ?
- The improvement in speedup and latency that we can achieve if we add a processor or two to a multiprocessor system ?
- The improvement in mean response time of a network if the copper wires are replaced by optical fiber ?

so

Terminology (1)

What is “system” in the context of performance evaluation ?

An assemblage of objects so combined by nature or human as to form an integral unit

A regularly interacting or interdependent group of objects forming a unified whole

Webster’s Dictionary

A combination of components/objects that act together to perform a function not possible with any of the individual parts

IEEE Standard Dictionary of Electrical and Electronic Terms

Two major features

1. A system consists of interacting objects/components
2. A system is associated with a function/work that it performs

Terminology (2)

Performance

Which questions are we thinking about when using the word “**performance**”?

Terminology (2)

Performance

Which questions are we thinking about when using the word "**performance**"?

- Basic concepts of work (performance metric):
  - Latency (time)
  - Bandwidth (rate)

Terminology (2)

Performance

Which questions are we thinking about when using the word "**performance**"?

- Basic concepts of work (performance metric):
  - Latency (time)
  - Bandwidth (rate)
- How well a computer system performs a given job or activity (workload)?

Terminology (2)

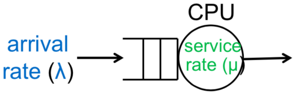
Performance

Which questions are we thinking about when using the word "**performance**"?

- Basic concepts of work (performance metric):
  - Latency (time)
  - Bandwidth (rate)
- How well a computer system performs a given job or activity (workload)?
- What is hard?
  - Performance of a computer system is multidimensional
    Complex component interaction; hard to predict how it will scale;...

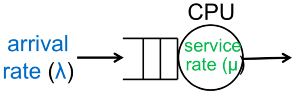
Terminology (2)

Questions on performance (in the course)



Terminology (2)

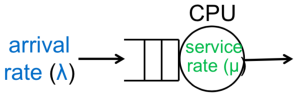
Questions on performance (in the course)



- What is the average time it takes a job to complete service?
- What is the throughput of the system (number of jobs completed per unit time)?
- If arrival rate is doubled ($\lambda \to 2\lambda$), how much should μ increase? Do we do nothing or do we need another CPU?

Terminology (2)

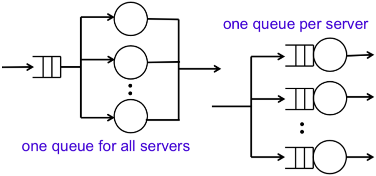
Questions on performance (in the course)



- What is the average time it takes a job to complete service?
- What is the throughput of the system (number of jobs completed per unit time)?
- If arrival rate is doubled ($\lambda \to 2\lambda$), how much should μ increase? Do we do nothing or do we need another CPU?

If we need more server capacity, one queue per server what are our options?

- buy a new server with the needed capacity
- buy a few smaller servers that adds up to the required capacity
  1. one queue for all servers
  2. one queue for each server



In which courses performance discussed?

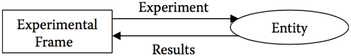
- Digital systems
- Logic design with HDL
- Data structures and algorithms
- Computer architectures
- Operating systems
- Microprocessors
- Embedded systems
- Computer networks
- Distributed systems
- Management for engineers
- ...
- Graduation project

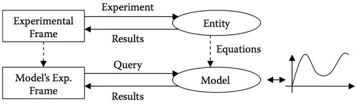
Performance evaluation: HOW

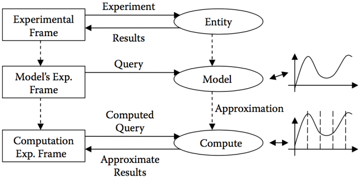
- Measurements of actual systems
- Simulations using software systems
- Mathematical modeling using techniques as queuing analysis

Performance evaluation: HOW

- Measurements of actual systems
- Simulations using software systems
- Mathematical modeling using techniques as queuing analysis

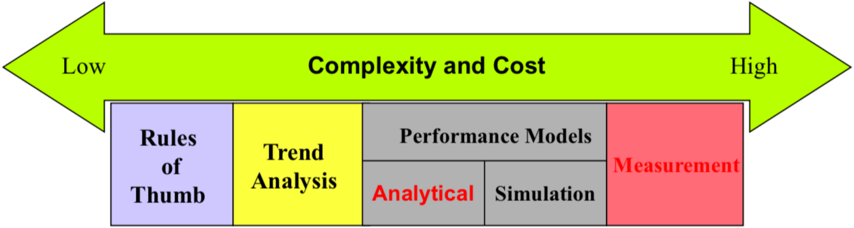






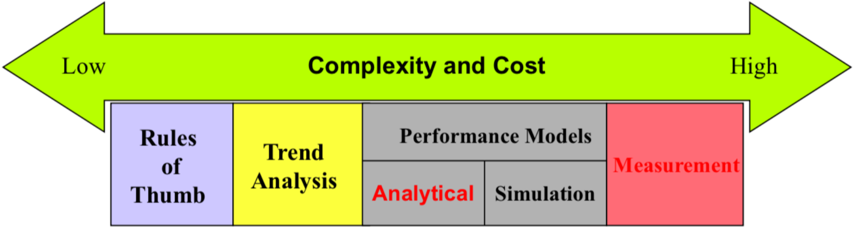
Performance evaluation: HOW

Complexity and Cost



Performance evaluation: HOW

Complexity and Cost



Which methods we (as students) often use in practice?

Goals of performance evaluation

- **Compare alternative system designs**
  Example: How much memory for shortest path algorithms ?
- **Procurement**
  Example: finding a cost-effective database for a specific application ?
- **Capacity planning**
  Example: Using available resources for optimum performance
- **System tunning**
  Example: finding best set of parameters for a high performance system
- **Performance debugging**
  Example: a system runs slowly not as expected in design. Trying to detect reasons for that
- **Set expectation/Recognize relative performance**
  Example: to “guess” something not jet happen, but be useful for planning.

Outline

1 Motivating examples

2 Basic concepts
- Performance of a system
- How to know system performance?

3 Course overview

4 Case-study
- Supply chain by queueing network
- Shortest path algorithms
- “Art” in performance evaluation

Main goals of the course

The course has two main goals

- Introducing a broad knowledge on performance evaluation and its application in computer science and engineering
- Presenting basic steps in performance evaluation, including
  - Performance metrics and workload
  - Analytical modeling
  - Experimental design
  - Simulation

Objectives/outcomes

- Specifying performance requirements
- Characterizing the load on the system (workload characterization)
- Finding the performance bottleneck (bottleneck identification)
- Determining the number and sizes of components (capacity planning)
- Evaluating design alternatives

Course evaluation

Part	**Percentage**	Assessment method
Quiz 1 (Kiểm tra 1)	20	Multiple choice (required)
Quiz 2 (Kiểm tra 2)		Multiple choice
Quiz 3 (Kiểm tra 3)		Multiple choice
Project 1 (Bài tập lớn 1)	30	Analytical modeling & presentation (required)
Project 2 (Bài tập lớn 2)		Python programming & presentation
Final exam (Thi cuối kỳ)	50	Multiple choice, short answer (required)

Course evaluation

Part	Percentage	Assessment method
Quiz 1 (Kiểm tra 1)	20	Multiple choice (required)
Quiz 2 (Kiểm tra 2)		Multiple choice
Quiz 3 (Kiểm tra 3)		Multiple choice
Project 1 (Bài tập lớn 1)	30	Analytical modeling & presentation (required)
Project 2 (Bài tập lớn 2)		Python programming & presentation
Final exam (Thi cuối kỳ)	50	Multiple choice, short answer (required)

Changes in assessment plan

The assessment plan is tentative.

Studying methods suggested

- Study the glossary of symbols and definitions to get to know the “language”.
- Read ahead material before each lecture.
- Do lots of exercies, look at problems early from the begining of the semester.
- Redo your homework until it is correct.
- To be encouraged active learning.

References

- **“Art of Computer Systems Performance Analysis”, R. Jain, Wiley, 1991, ISBN:0471503363** (Winner of the “1992 Best Computer Systems Book” Award from Computer Press Association”)
- “Computer Systems Performance Evaluation and Prediction”, Paul Fortier and Howard Michel, Digital Press, 2000.
- “Fundamentals of Performance Evaluation of Computer and Telecommunication Systems”, Mohammad S. Obaidat and Noureddine A. Boudriga, Wiley, 2010.
- Internet

Outline

1 Motivating examples

2 Basic concepts

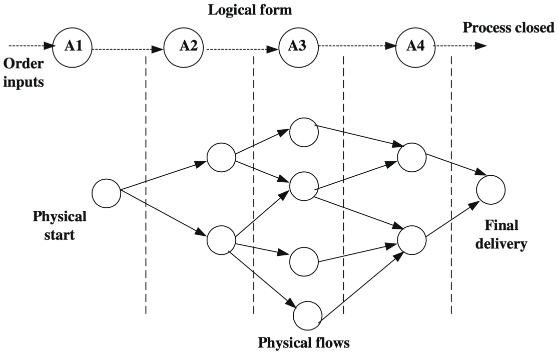
- Performance of a system
- How to know system performance?

3 Course overview

4 Case-study

- Supply chain by queueing network
- Shortest path algorithms
- “Art” in performance evaluation

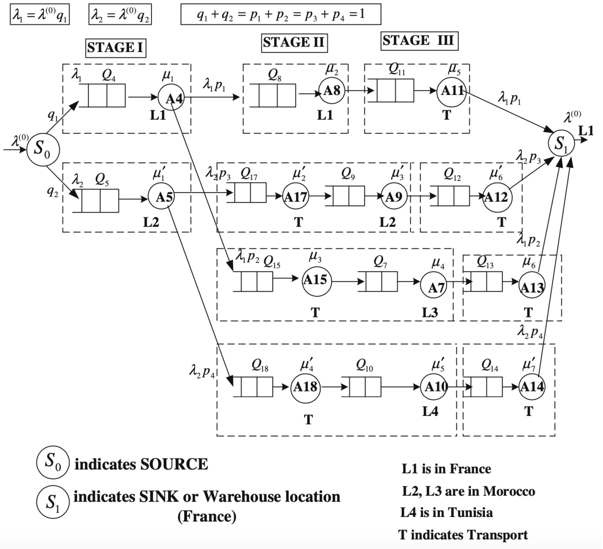
Problem and objectives



Objectives

- Minimum response time (to deliver an order to the final destination.
- Average number of items delivered with this minimum response time

Supply chain: queueing network



indicates SOURCE

indicates SINK or Warehouse location (France)

L1 is in France

L2, L3 are in Morocco

L4 is in Tunisia

T indicates Transport

Algorithm: shortest path algorithms

Reminder

- Which algorithms have you learnt ?
- What do you remember on them in terms of characteristics ?
- How to know which one is the best ? How to compare among them ?

Algorithm: shortest path algorithms

Reminder

- Which algorithms have you learnt ?
- What do you remember on them in terms of characteristics ?
- How to know which one is the best ? How to compare among them ?

Problem statement

Given a large sparse weighted graph. How to evaluate the performance of shortest path algorithms ?

Shortest path problems

- How large in number of vertices $|V|$?

Shortest path problems

- How large in number of vertices $|V|$?
- How sparse in vertex degree ?

Shortest path problems

- How large in number of vertices $|V|$?
- How sparse in vertex degree ?
- Weight function: uniform distribution in a given range $[a, b]$

Shortest path problems

- How large in number of vertices $|V|$?
- How sparse in vertex degree ?
- Weight function: uniform distribution in a given range $[a, b]$
- Source vertex: full sampling (i.e., all vertices)

Shortest path problems

- How large in number of vertices $|V|$?
- How sparse in vertex degree ?
- Weight function: uniform distribution in a given range $[a, b]$
- Source vertex: full sampling (i.e., all vertices)

Performance metrics

Shortest path problems

- How large in number of vertices $|V|$?
- How sparse in vertex degree ?
- Weight function: uniform distribution in a given range $[a, b]$
- Source vertex: full sampling (i.e., all vertices)

Performance metrics

- running time

Shortest path problems

- How large in number of vertices $|V|$?
- How sparse in vertex degree ?
- Weight function: uniform distribution in a given range $[a, b]$
- Source vertex: full sampling (i.e., all vertices)

Performance metrics

- running time
- memory consumption

Shortest path problems

- How large in number of vertices $|V|$?
- How sparse in vertex degree ?
- Weight function: uniform distribution in a given range $[a, b]$
- Source vertex: full sampling (i.e., all vertices)

Performance metrics

- running time
- memory consumption
- Time to shortest path solution

Shortest path problems

- How large in number of vertices $|V|$?
- How sparse in vertex degree ?
- Weight function: uniform distribution in a given range $[a, b]$
- Source vertex: full sampling (i.e., all vertices)

Performance metrics

- running time
- memory consumption
- Time to shortest path solution
- How many times to swap memory in/out (for limited memory case) ?

Shortest path problems

- How large in number of vertices $|V|$?
- How sparse in vertex degree ?
- Weight function: uniform distribution in a given range $[a, b]$
- Source vertex: full sampling (i.e., all vertices)

Performance metrics

- running time
- memory consumption
- Time to shortest path solution
- How many times to swap memory in/out (for limited memory case) ?
- ...

Shortest path problems

- How large in number of vertices $|V|$?
- How sparse in vertex degree ?
- Weight function: uniform distribution in a given range $[a, b]$
- Source vertex: full sampling (i.e., all vertices)

Performance metrics

- running time

Challenge

Test performance of different one source-all destinations shortest path algorithms (i.e., Dijkstra, Bellman-Ford)
all parameters and implementation variants should be considered

Discussion

What happens (for shortest path algorithms/implementations) if computer memory cannot store the data structure of the whole graph ?

How to compare experimental results

Prof. Raj Jain

Performance evaluation is an art.

Throughputs (transactions per second) of 2 systems A and B

System	Workload 1	Workload 2
A	20	10
B	10	20

How to compare experimental results

Prof. Raj Jain

Performance evaluation is an art.

Throughputs (transactions per second) of 2 systems A and B

System	Workload 1	Workload 2
A	20	10
B	10	20

There are 3 possible ways to compare (ratio game).

System	Average	Average(*/B)	Average(*/A)
A	15	(2+0.5/2)=1.25	(1+1)/2=1
B	15	(1+1)/2=1	(0.5+2)/2=1.25

Enjoy your study :-)