Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный университет телекоммуникаций и информатики»

Кафедра вычислительных систем

Расчетно-графическое задание по дисциплине

«Архитектура вычислительных систем»

Вариант 15

Выполнил:

Студент гр. ИП-813

Бурдуковский И.А.

Проверил:

к.т.н доцент кафедры ВС

Ефимов А.В.

Новосибирск 2020

ОГЛАВЛЕНИЕ

[Часть 1 3](#_Toc59456391)

[1.1 Задание 3](#_Toc59456392)

[1.2 Общие сведения 3](#_Toc59456393)

[1.3 Характеристики 4](#_Toc59456394)

[1.4 Иерархия структур коммуникационных сетей 5](#_Toc59456395)

[1.5 Технологии программирования 13](#_Toc59456396)

[1.6 Вывод 15](#_Toc59456397)

[Часть 2 16](#_Toc59456398)

[2.1 Задание 16](#_Toc59456399)

[2.2 Расчёт диаметра 16](#_Toc59456400)

[2.3 Расчёт среднего диаметра 17](#_Toc59456401)

[2.4 Расчёт бисекционной пропускной способности 17](#_Toc59456402)

[Список использованной литературы 18](#_Toc59456403)

# Задание 1

## Задание

Выполнить анализ архитектуры суперВС PANGEA III (№ 18 в списке Top500 по данным ноября 2020).

## Общие сведения

PANGEA III — это суперкомпьютер, установленный в Центре технических и научных исследований CSTJF в Пау, Франция. Он был запущен в июне 2019 и сразу появился в списке top500 на 11ой позиции. С тех пор начал сдавать позиции – 15ое, а затем и 18ое место к текущему моменту времени.

PANGEA III похож по архитектуре на своих старших братьев, произведённых IBM: суперкомпьютеры Siera и [Summit](https://ru.wikipedia.org/wiki/Summit_(%D1%81%D1%83%D0%BF%D0%B5%D1%80%D0%BA%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80)), но производительность меньше: теоретическая производительность 25 petaFLOPS против 125 и 200 petaFLOPS соответственно.



Рис.1. Суперкомпьютерный комплекс PANGEA III

PANGEA III используется для обработки данных на этапе разведки и разработки недр, чтобы можно было точно определять местонахождение углеводородов под землей. Кроме того, компания Total заявляет о внедрении новых методов пластового моделирования, чтобы лучше прогнозировать объём добычи ресурсов.

## Характеристики

PANGEA III имеет следующие технические характеристики и программное обеспечение:

* Процессоры: IBM POWER9 18C 3.45ГГц
* Графические процессоры: NVIDIA Volta GV100
* Узлов входа: неизвестно
* Всего узлов: неизвестно
* Ядер на узел: 36
* Всего ядер: 291 024
* Количество CPU: неизвестно
* Общий объём RAM: 288 768 GB
* Тип RAM: DDR4 2666 MHZ
* Память процессора: 256 GB на узел
* Количество GPU: неизвестно
* Память GPU: 64 ГБ на узел
* Пиковая пропускная способность памяти на узел: 340 ГБ/сек.
* Операционная система: Red Hat Enterprise Linux
* Параллельная файловая система: IBM Spectrum MPI
* Пакетная система: IBM Spectrum LSF
* Сеть внутренних коммутаций: IB Enhanced Data Rate
* Пиковая производительность: 25 025 TFLOPS

## Иерархия структур коммуникационных сетей

Начнем рассматривать суперкомпьютер PANGEA III c наименьшей единицы – ядра.

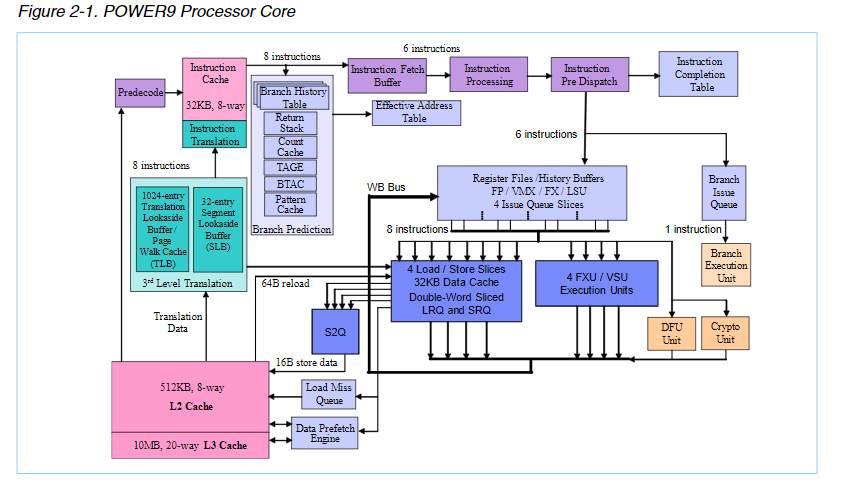


Рис.2. Структура процессорного ядра

POWER 9 вносит много усовершенствований ядра, а также большие архитектурные изменения со времён POWER 8. IBM использует высокомодульный подход к дизайну и предлагает POWER 9 в двух различных вариантах исполнения: Scale Out и Scale Up.

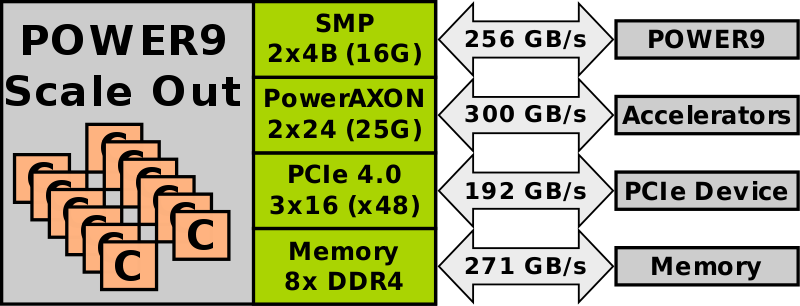


Рис.3. 12-ядерная модель Scale Out

* Scale-Out:
  + Предназначен для традиционных кластеров центров обработки данных, использующих серверы с одним сокетом и двумя сокетами.
  + Оптимизирован для серверов Linux
  + Представлены в двух вариациях: 12-ядерные и 24-ядерные модели
  + Предлагают пропускную способность вплоть до 120 GB/s

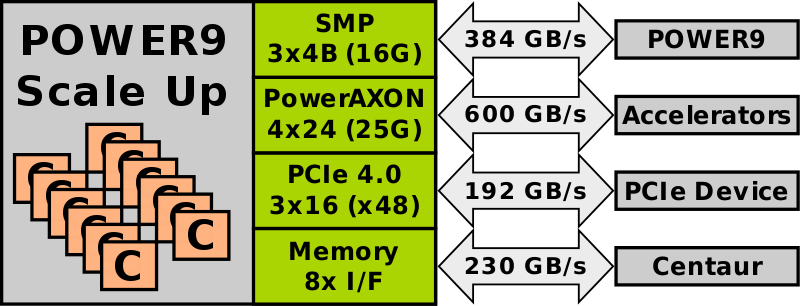


Рис.4. 12-ядерная модель Scale Up

* Scale-Up:
  + Предназначен для корпоративных серверов NUMA с четырьмя и более сокетами, поддерживающими большой объем памяти и пропускную способность.
  + Оптимизирован для серверов PowerVM
  + Представлены в двух вариациях: 12-ядерные и 24-ядерные модели

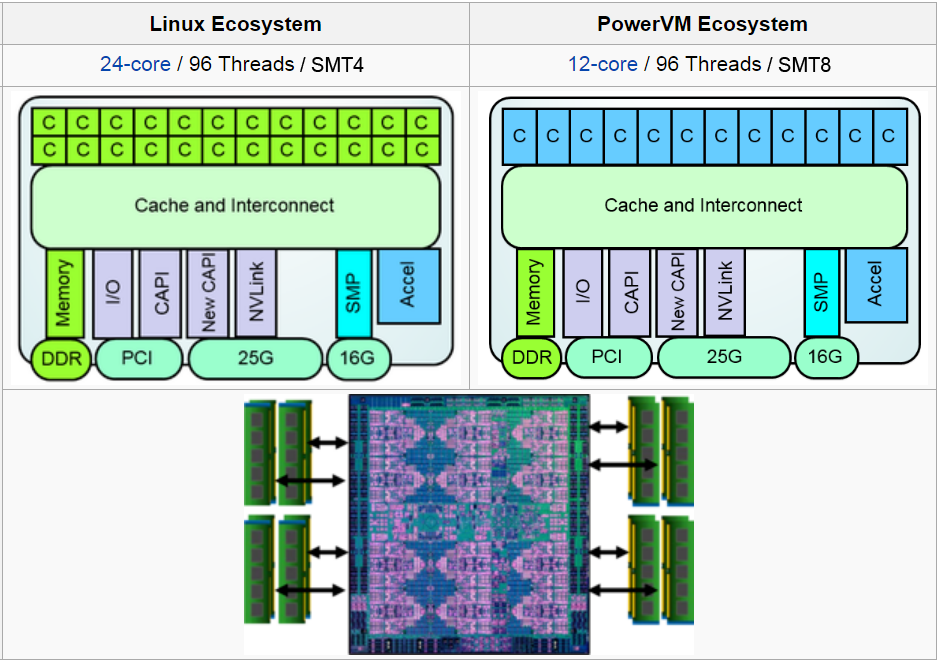


Рис.5. 12-ядерная и 24-ядерная Scale Out модели

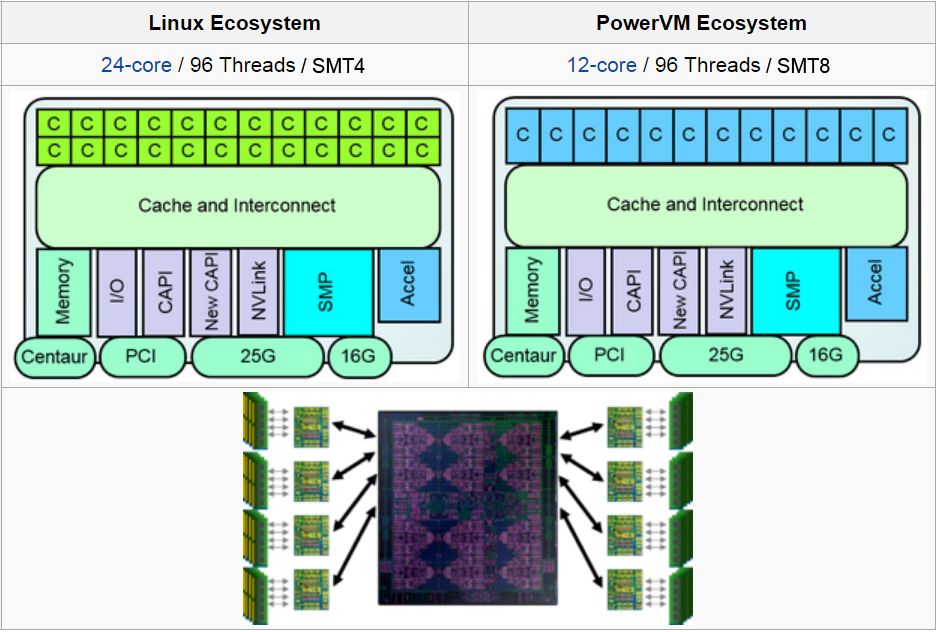


Рис.6. 12-ядерная и 24-ядерная Scale Up модели

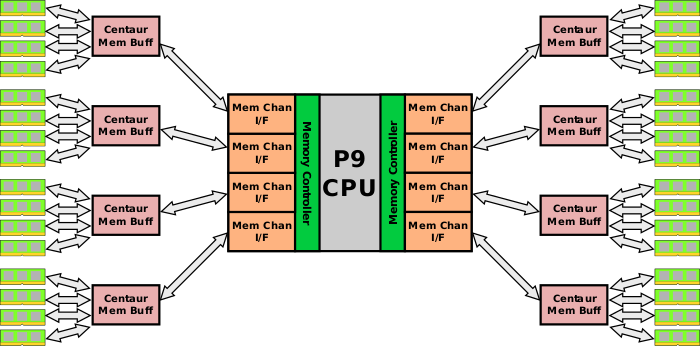


Рис.7. Архитектура буферной памяти Scale Up модели

POWER 9 имеет 2 контроллера памяти, подходящими для работы с 4-мя

DMI каналами. Каждый DMI канал подключается к одному отдельному чипу буферной памяти “Centaur”, который обеспечивает 4 канала памяти DDR4

POWER9 может иметь переменное число активных ядер: 16, 18, 20, 22. Каждый процессор POWER9 суперкомпьютера PANGEA III содержит 18 ядер. Благодаря встроенному коммутатору, соединяющему ядра с памятью, графическими процессорами и остальным, совокупная пропускная способность высокопроизводительной матрицы на кристалле 7 ТБ/с.

Технологический процесс составляет 14 нм.

●      кэш:

○      данных L1: 32 KB на ядро, 8-way associative

○      инструкций L1: 32 KB на ядро, 8-way associative

○      L2: 512 KB на ядро, 8-way associative

○      L3: 120 MB, 20-way, 10 MB/core pair

Частота: благодаря адаптивным опциям управления питанием тактовая частота может варьироваться в зависимости от нагрузки системы. Скорости могут варьироваться примерно от 3,15 до 3,45 ГГц.

Архитектура Узла:

Каждый узел IBM POWER9 AC922 имеет два процессорных сокета IBM POWER9 и поддерживает до 6 графических видеокарт NVIDIA Volta V100. По умолчанию встречаются вариации с 4-мя и 6-тью графическами процессорами. Какой из вариантов был выбран для PANGEA III – неизвестно. В качестве интерконнекта для CPU-GPU и GPU-GPU используется технология NVLink2.

Узел поддерживает суммарно до 16 ТБ основной (оперативной) памяти, для её размещения имеется 16 сокетов DDR4, по 8 DIMM DDR4 на процессорный сокет. Вычислительные узлы PANGEA III имеют по 256 ГБ основной памяти. Пропускная способность канала CPU-RAM составляет 170 ГБ/с на процессор, или 340 ГБ/с на узел.

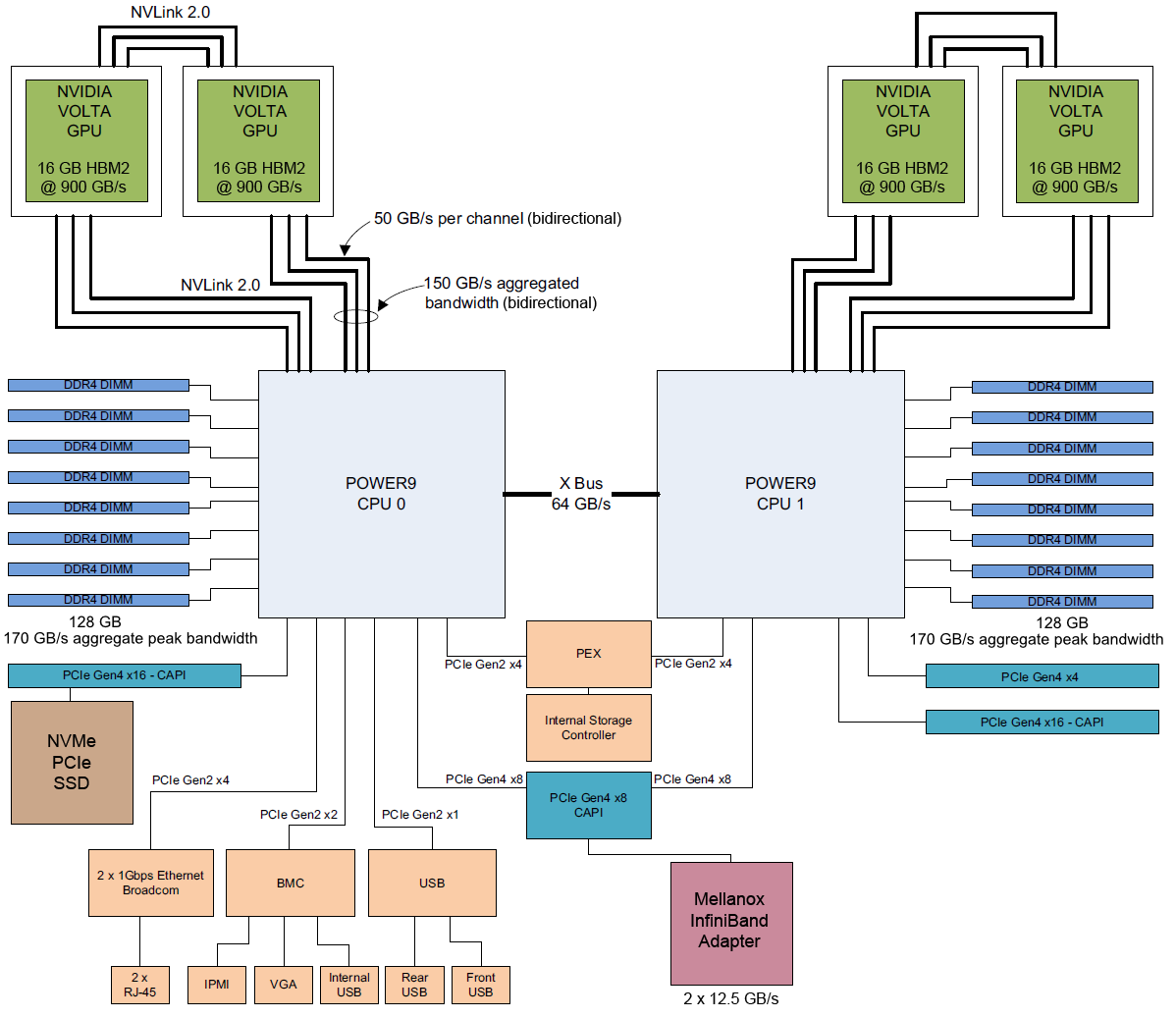


Рис.8. Диаграмма узла POWER9 с 4-мя графическими процессорами

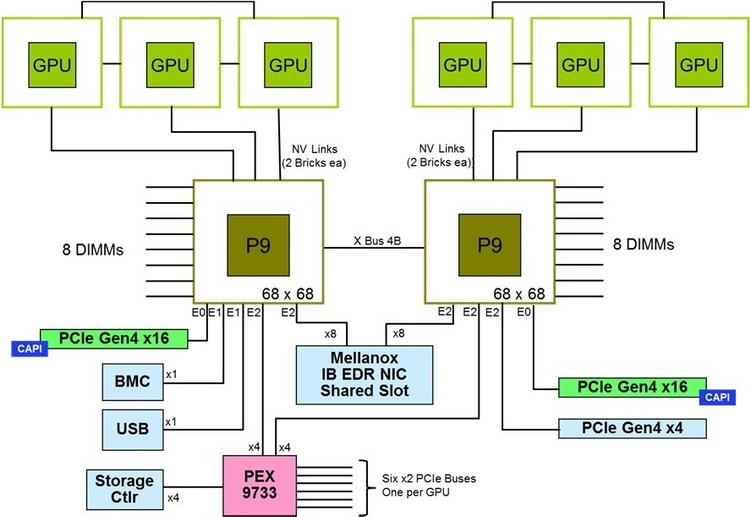


Рис.9. Диаграмма узла POWER9 с 6-тью графическими процессорами

Технология NVLink 2.0 поддерживает до 6 NVLink ссылок на один графический процессор. Каждая ссылка обеспечивает двунаправленное соединение со скоростью 50 ГБ/с к другому графическому процессору или центральному процессору, что обеспечивает суммарную пропускную способность 300 ГБ/с.

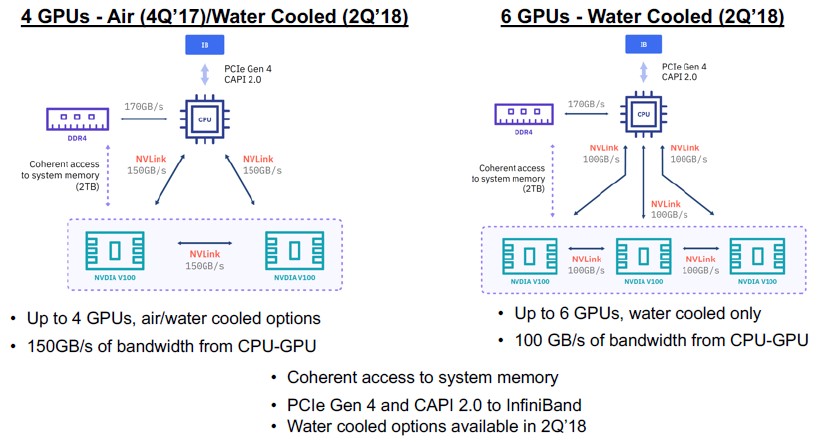


Рис.10. V100 с подключенным по NVLink GPU-GPU и GPU-CPU (2 и 3 GPU)

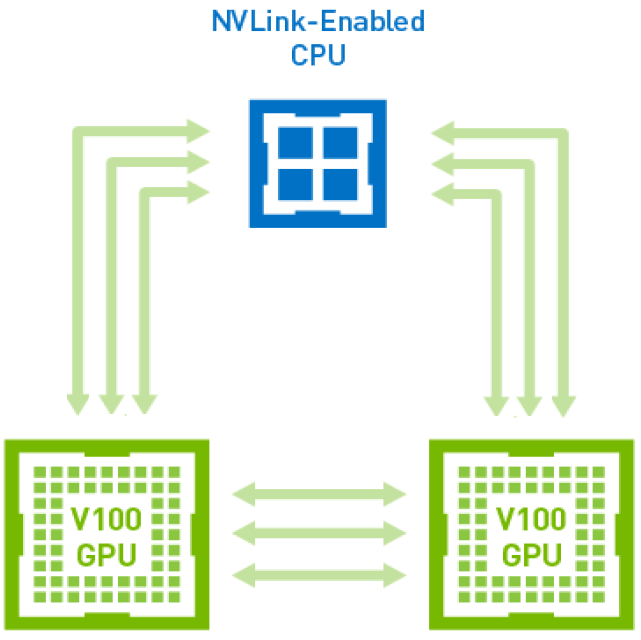


Рис.11. V100 с подключенным по NVLink GPU-GPU и GPU-CPU

Каждый ЦП подключен к 2-м или 3-м графическим процессорам по 3 NVLink в каждом. Эти графические процессоры соединены друг с другом 3 NVLinks каждый. Но графические процессоры на разных процессорах не соединяются друг с другом с помощью NVLinks.

Центральные процессоры соединяются с контроллерами, устройствами ввода и вывода и пр. с помощью шины PCI Express 4.0 (PCIe) со скоростью обмена до 16 Гбит/с на линию. Устройства PCI Express взаимодействуют между собой через среду, образованную коммутаторами, при этом каждое устройство напрямую связано соединением типа точка-точка с коммутатором.

Два центральных процессор, до шести графических процессоров, диски (2 отсека для дисков: для 2 жестких дисков (HDD) или 2 твердотельных дисков (SSD)) и прочее образуют узел кластера.

Узел является основным блоком кластера Linux, который, по сути, является автономным, бездисковым, многоядерным независимым компьютером, с возможностью монтажа в серверную стойку и удалённым управлением.

Узлы являются частью «Scalable Unit» (масштабируемая единица) или же сокращенно SU. Кроме узлов туда входят коммутаторы, разное оборудование управления, рамки, достаточные для размещения всего необходимого оборудования.

Связь между узлами происходит с помощью коммутаторов первого уровня. А связь между несколькими SU происходит с помощью коммутаторов второго уровня.

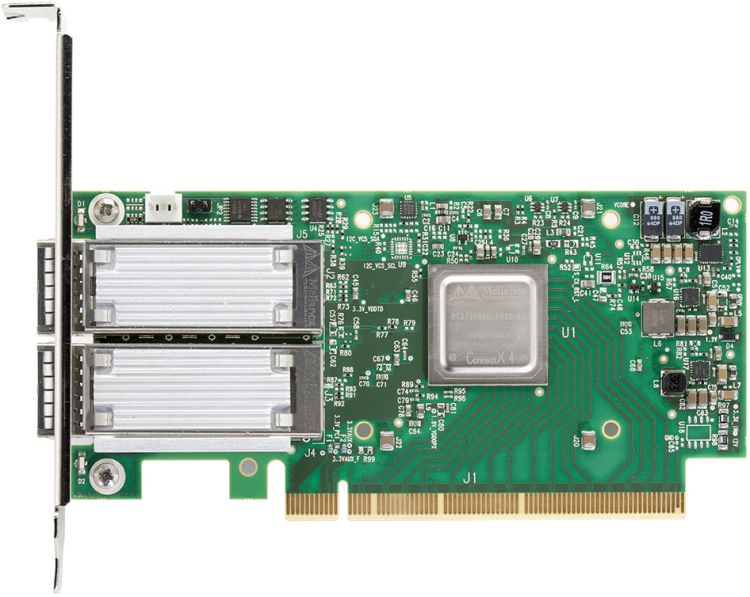


Рис.12.Адаптер Mellanox EDR InfiniBand



Рис.13.Mellanox Switch-IB Top-of-Rack

Узлы имеют один двухпортовый адаптер Mellanox EDR Infiniband с передачей данных 100Гбит/с. Порты адаптера подключаются к коммутаторам первого уровня Mellanox Switch-IB с 36 портами, которые через оптоволоконные кабели подключаются к коммутаторам второго уровня Mellanox CS7500 с 648 портами.

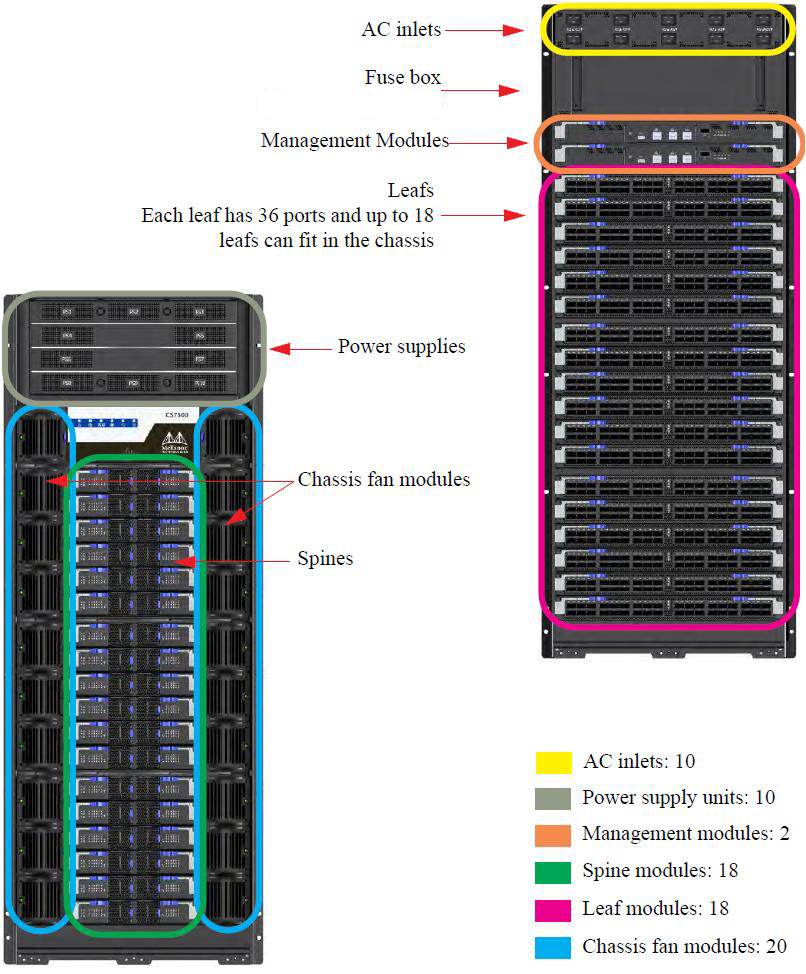


Рис. 14. Mellanox CS7500 с маркировкой

Коммутаторы образуют иерархию с коммутаторами более высокого уровня, имеющие большее количество соединений, чем коммутаторы более низкого уровня. Число отключений для переключателей нижнего уровня увеличивается в соотношении два к одному. Все узлы соединяются в единую сеть толстого дерева.

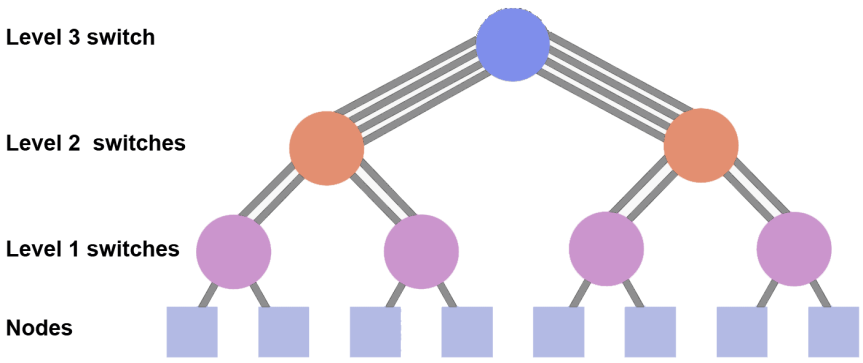


Рис. 15. Толстое дерево сети

Наверняка толстое дерево сети PANGEA III неизвестно из-за закрытости информации, но можно предположить, что оно имеет такой же вид как у его старших братьев Lassen и Siera.

Каждая стойка PANGEA III тогда имеет 18 узлов и 2 переключателя TOR. Двухпортовый адаптер каждого узла подключается к обоим коммутаторам TOR с одним портом каждый. Каждый коммутатор TOR имеет 12 каналов связи с коммутаторами Director, по одному на коммутатор Director. Существует девять переключателей Director. Поскольку каждый коммутатор TOR имеет 12 восходящих каналов и имеется только 9 коммутаторов Director, то на каждый коммутатор TOR существует ещё 3 дополнительных восходящих канала. Они используются для подключения к 3 из 9 коммутаторов Director.

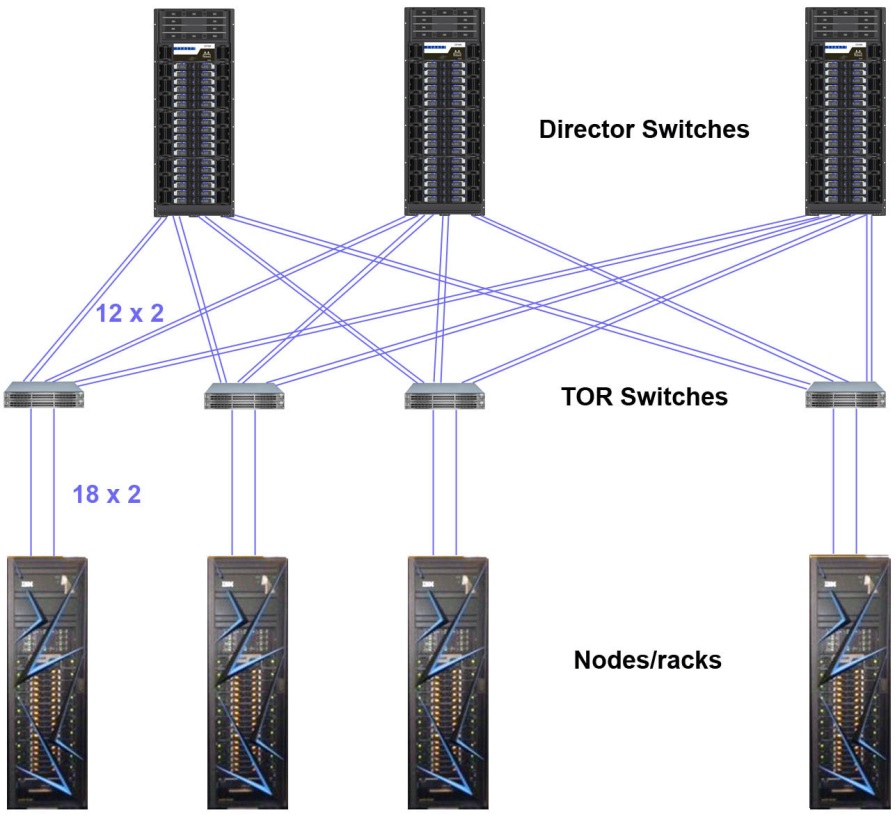


Рис. 16. Сеть Sierra (предполагаем что и PANGEA III)

На картинке видно только 18 восходящих ссылок на TOR, вместо 24 => толстое дерево Siera – модифицировано к виду 1.5:1.

Для измерения пропускной способности оборудования, измерили полосу пропусканий между узлами. Измерения были выполнены на работающих с высокой нагрузкой машинах с использованием простого неблокирующего тестового кода MPI. Одна задача на каждом из двух узлов, размер сообщения 1 МБ. В результате, произошла задержка равная одной наносекунде, а пропускная способность оказалась примерно равна 21 ГБ/с.

PANGEA III имеет структуру толстого дерева. В нем ветви ближе к вершине иерархии «толще», чем ветви ниже. В коммуникационной сети ветви являются каналами передачи данных; различная толщина (пропускная способность) каналов передачи данных обеспечивает более эффективное и специфичное для технологии использование, а также ускоряет работу системы и обеспечивает большую надежность.

Сеть толстых деревьев особенно хорошо подходит для вычислений с быстрым преобразованием Фурье, которые нужны для таких задач обработки сигналов, как радар, сонар, что как раз-таки может быть использовано в горно-добывающей промышленности, которой занимается компания Total, владеющая суперкомпьютером PANGEA III.

## Программная архитектура

Программное окружение, предоставляемое пользователям суперкомпьютерной системы PANGEA III, обладает следующими характеристиками:

● ОС: Red Hat Enterprise Linux — [дистрибутив](https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D1%82%D1%80%D0%B8%D0%B1%D1%83%D1%82%D0%B8%D0%B2_%D0%BE%D0%BF%D0%B5%D1%80%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D0%BE%D0%B9_%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D1%8B) [Linux](https://ru.wikipedia.org/wiki/Linux) компании [Red Hat](https://ru.wikipedia.org/wiki/Red_Hat), позиционирующийся для корпоративного использования. Новые версии выходят с периодичностью около 3 лет. Основная особенность дистрибутива — наличие коммерческой поддержки на протяжении 10 лет, с возможностью продления до 13 лет.

● Компиляторы:

○ IBM XLC

○ Clang: clang/3.9.0, clang/3.9.1, clang/4.0.0

○ GCC: gcc/4.8-redhat, gcc/4.9.3, gcc/6.1.0

○ Intel: intel/14.0.3, intel/16.0.3, intel/17.0.2, intel/18.0.1 (L,D), intel/15.0.6, intel/16.0.4, intel/18.0-beta, intel/18.0.2, intel/16.0.2, intel/17.0.0, intel/18.0.0

○ PGI: pgi/16.3, pgi/16.7, pgi/16.9

Математические библиотеки: ESSL, CUBLAS 9.2

MPI: IBM Spectrum MPI

## Вывод

В результате исследования я частично выяснил, особенности устройства суперкомпьютера PANGEA III. К сожалению, из-за закрытости информации о суперкомпьютере, в силу его коммерческого использования, не удалось выяснить ни количество узлов, ни выбора точного типа узла модели IBM POWER SYSTEM AC922, ни достоверного выбора связей между узлами. Было принято решение допускать схожесть устройства суперВС с её близкими родственниками от одного производителя IBM. Благо намёков на правоту этих допущений достаточно из интервью с самим производителем.

Ядра процессора соединяются посредствам высокопроизводительной матрицы на кристалле с пропускной способностью 7 ТБ/с. Графические процессоры связаны между собой и с центральным процессором с помощью технологии NVLink 2.0,, которая поддерживает до 6 соединений и суммарную пропускную способность 300 ГБ/с. Центральный процессор соединён с другими составляющими через шину PCIe со скоростью обмена до 16 Гбит/с на линию. Все вышеперечисленное объединяется в узлы.

Узлы связываются в единую сеть посредством топологии Толстого дерева. Связь между узлами происходит с помощью коммутаторов первого уровня Mellanox Switch-IB с 36 портами. А связь между несколькими SU, куда входят узлы, происходит с помощью коммутаторов второго уровня Mellanox CS7500 с 648 портами.

Подводя итоги, можно сказать, что суперВС PANGEA III (18 место в топ 500), принадлежащий компании Total, является уменьшенным подобием суперкомпьютеров Lassen, Sierra и Summit, разделившие 17, 3 и 2 места соответственно. Сейчас от Lassen’a PANGEA III отстаёт лишь на 1 место. При желании, благодаря особенности топологии толстого дерева, суперкомпьютер можно увеличить чтобы достичь производительности своих конкурентов.

# Задание 2

## Задание:

Расчёт структурных характеристик:

* Диаметр структуры
* Средний диаметра
* Бисcекционная пропускная способности

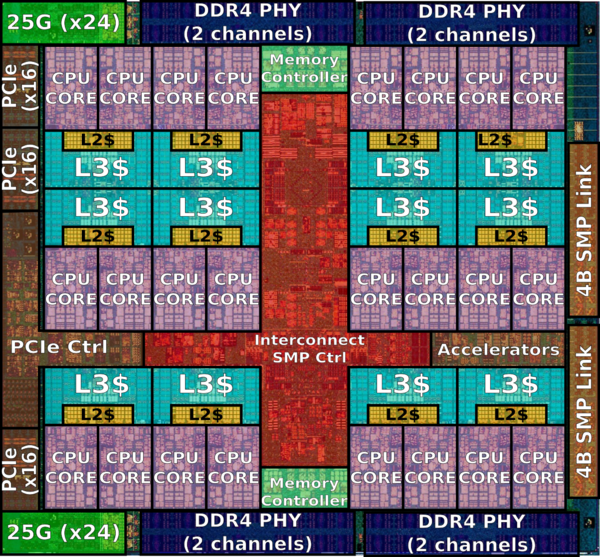
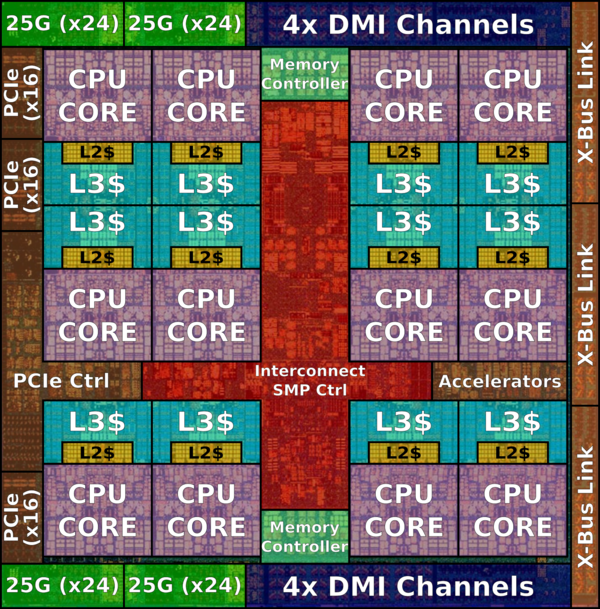


Рис. 17. SMP Архитектура процессоров POWER 9 c 12-тью и 24-мя ядрами

Имея типовые данные можем получить:

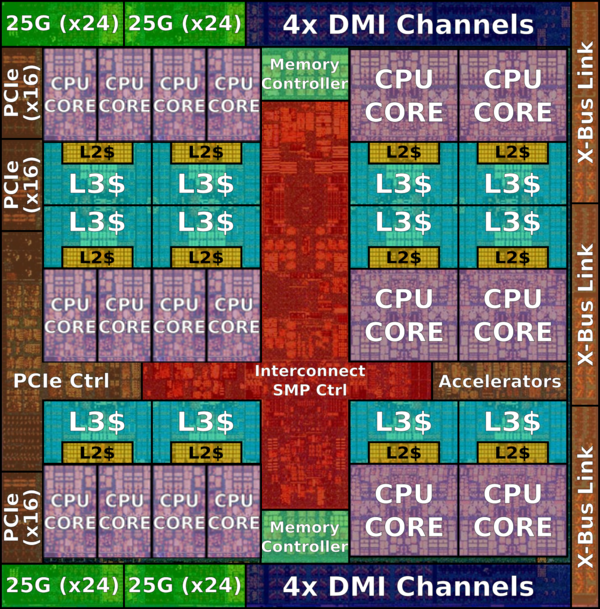


Рис. 18. SMP Архитектура процессора POWER 9 c 18-тью ядрами

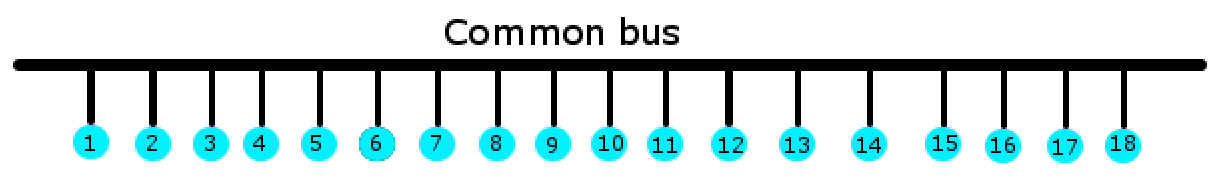


Рис. 19. Условный граф SMP архитектуры IBM POWER 9 18C

1. Расчёт диаметра:

Диаметр структуры - максимальное расстояние, определённое на множестве кратчайших путей между парами вершин структуры ВС

𝑑= {}

*– расстояние, т.е. минимальное число рёбер, образующих путь из вершины i в вершину j ; i, j ∈ {0, 1, ..., N −1};*

Расстояние между вершинами = 2 (1⇒Общая шина⇒2)

1. Расчёт среднего диаметра

*– число вершин, находящихся на расстоянии от любой выделенной вершины (однородного) графа G.*

*N – количество вершин*

Ответ: = 1,94

## Расчёт биссекционной пропускной способности

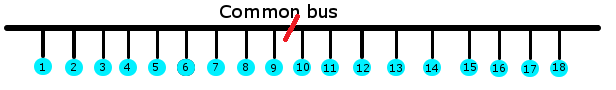


Рис. 20. Разделение графа на 2 части

Т.к. после разреза на 2 части, получили 1 разрыв, то биссекционная пропускная способность =1

# Список использованной литературы

* + 1. TOP500 the list [Электронный ресурс]. – URL: https://www.top500.org/system/179689/
    2. Total’s Pangea III Supercomputer Ranked First in Industry Worldwide [Электронный ресурс]. – URL: https://www.total.com/media/news/press-releases/totals-pangea-iii-supercomputer-ranked-first-industry-worldwide
    3. PANGEA: 80,000 TIMES MORE POWERFUL THAN YOUR AVERAGE PC [Электронный ресурс].– URL: https://www.total.com/media/news/news/pangea-3-times-more-powerful
    4. TADVISER Total Pangea III [Электронный ресурс].– URL: https://www.tadviser.ru/index.php/Продукт:Total\_Pangea\_III\_(суперкомпьютер)
    5. Pikabu статья про 12 суперкомпьютеров [Электронный ресурс]. – URL: https://pikabu.ru/story/12\_samyikh\_byistryikh\_superkompyuterov\_v\_mire\_v\_2020\_godu\_7808807
    6. IBM Develops World's Most Powerful Commercial Supercomputer for Total [Электронный ресурс]. – URL: https://newsroom.ibm.com/2019-06-18-IBM-Develops-Worlds-Most-Powerful-Commercial-Supercomputer-for-Total
    7. GEOExPro about Pangea III [Электронный ресурс]. – URL: https://www.geoexpro.com/articles/2020/03/pangea-iii-a-supercomputer-for-oil-and-gas-exploration
    8. Energy Giant takes the AI supercomputer route [Электронный ресурс]. – URL: https://www.nextplatform.com/2019/06/18/energy-giant-takes-the-ai-supercomputer-route/
    9. IBM Power System AC922 [Электронный ресурс]. – URL: https://www.ibm.com/downloads/cas/EPNDE9D0
    10. Microway IBM Power System AC922 [Электронный ресурс]. – URL: https://www.microway.com/product/ibm-power-systems-ac922/
    11. Power 9 Microarchitectures [Электронный ресурс]. – URL: https://en.wikichip.org/wiki/ibm/microarchitectures/power9
    12. IBM Power 9 broshure [Электронный ресурс]. – URL: https://www.ibm.com/downloads/cas/US-ENUS119-042-CA
    13. IBM Power 9 Processor User’s Manual [Электронный ресурс]. – URL: https://openpowerfoundation.org/?resource\_lib=power9-processor-users-manual
    14. POWER9 ИДЁТ В НАРОД: Описание сервера IBM AC922 [Электронный ресурс]. – URL: https://www.sekventa.ru/press-center/news/power9-idyot-v-narod-opisanie-servera-ibm-ac922.html
    15. Red Hat Enterprise Linux [Электронный ресурс]. – URL: https://ru.wikipedia.org/wiki/Red\_Hat\_Enterprise\_Linux
    16. Суперкомпьютер Summit [Электронный ресурс]. – URL: https://servernews.ru/961817
    17. Суперкомпьютеры IBM Summit и IBM Sierra [Электронный ресурс]. – URL: https://parallel.ru/computers/reviews/ibm\_summit\_sierra.html
    18. CS7500 InfiniBand Switch [Электронный ресурс]. – URL: <https://hpc.llnl.gov/sites/default/files/Mellanox.CS7500switch.pdf>
    19. Compilers at LC [Электронный ресурс]. – URL: https://hpc.llnl.gov/software/development-environment-software/compilers
    20. Using LC’s Sierra Systems [Электронный ресурс]. – URL: https://hpc.llnl.gov/training/tutorials/using-lcs-sierra-system#Hardware
    21. Livermore Computing Center [Электронный ресурс]. – URL: <https://hpc.llnl.gov/hardware/platforms/sierra>
    22. Ливерморская национальная лаборатория им. Лоуренса [Электронный ресурс]. – URL: https://hpc.llnl.gov/training/tutorials/livermore-computing-resources-and-environment