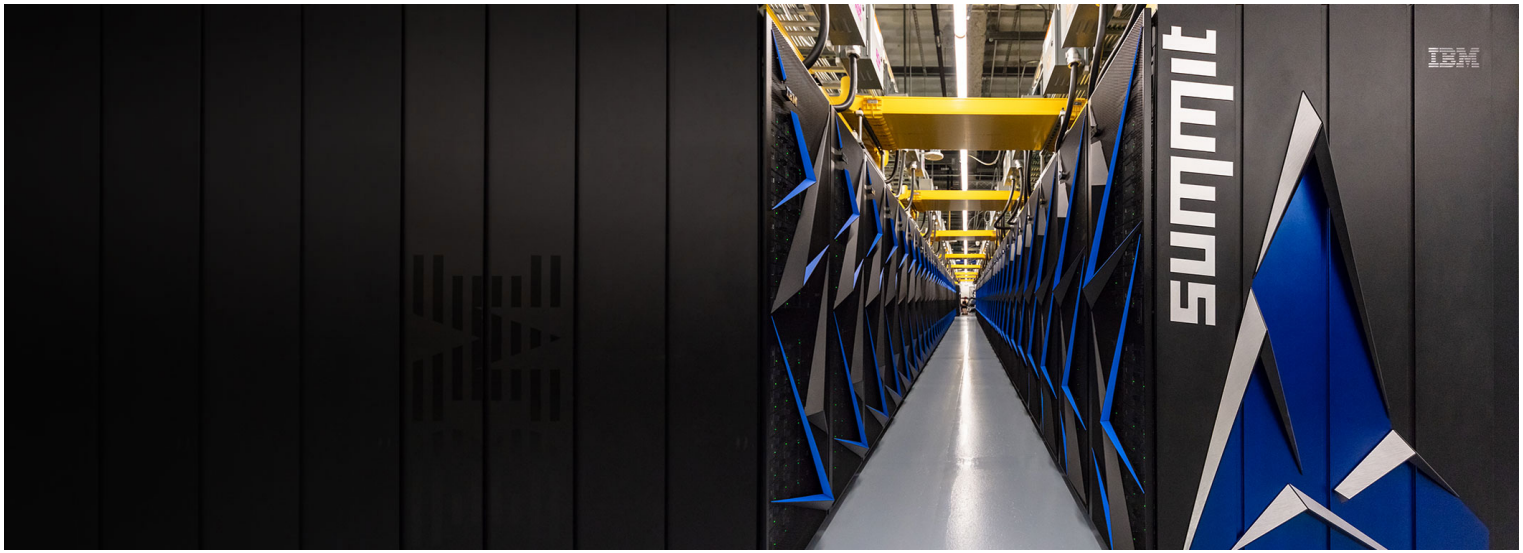
Коммуникационная среда в суперкомпьютерах

Алексей Николаевич Сальников

Факультет вычислительной математики и кибернетики
Московский государственный университет имени М.В. Ломоносова

17 марта 2020 г.

Суперкомпьютер "Summit" США – 2019



- Произв.: 148600 TFLOPS
- Число узлов: 4608 (на узел – процессоров: 2, ядер 22, GPU: 6)
- Операционная система: RedHat Linux 7.4
- оперативная память на узел: 512GB DDR4 + 96GB HBM2

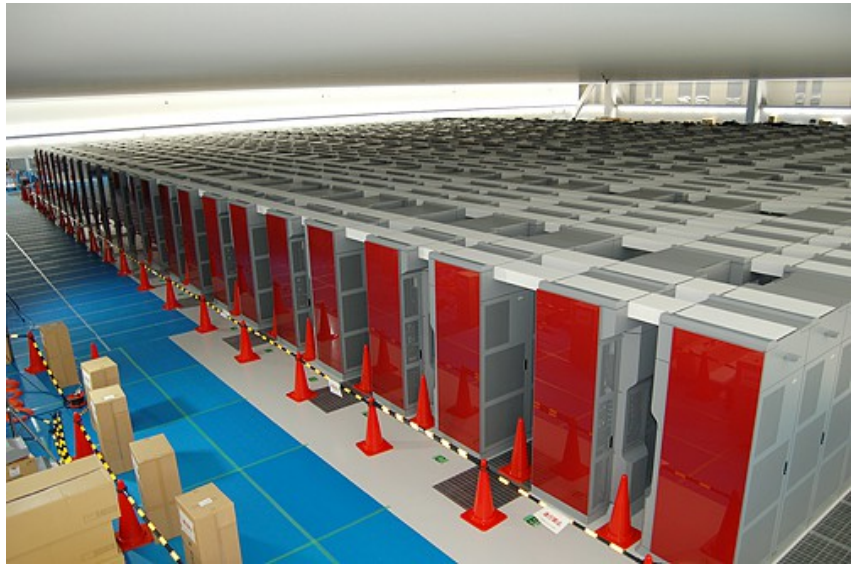
1-TFLOP - 10^{12} операций в секунду

Суперкомпьютер “Ломоносов” Россия – 2010

- Произв.: 397 TFLOPS
- Число узлов: 10260 (ядер 41040)
- Операционная система: Clustrx (модифицированный Linux)
- Объём оперативной памяти: 73920Gb
- Объём файлового хранилища: 1382400Gb

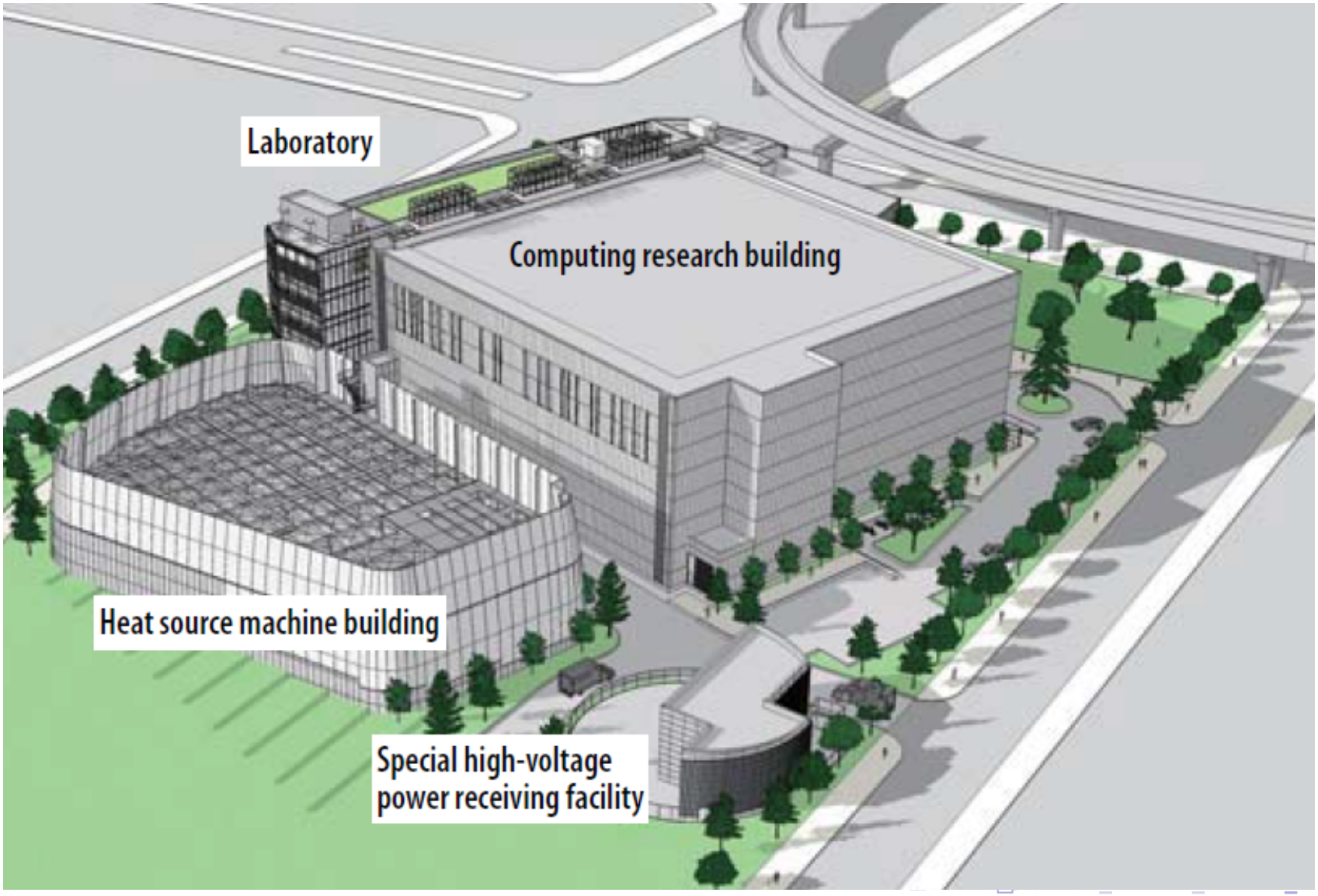
Суперкомпьютер “K” Япония - 2011

- Производительность: 10510 TFLOPS (10510000000000000 операций)
- Число процессоров: 88128 (ядер 705024)
- Операционная система: модифицированный Linux
- Объём оперативной памяти: 1410048 Gb



Суперкомпьютер “K” Япония - 2011

Структура вычислительного центра “Riken”



Список Top-500 (Ноябрь 2019)

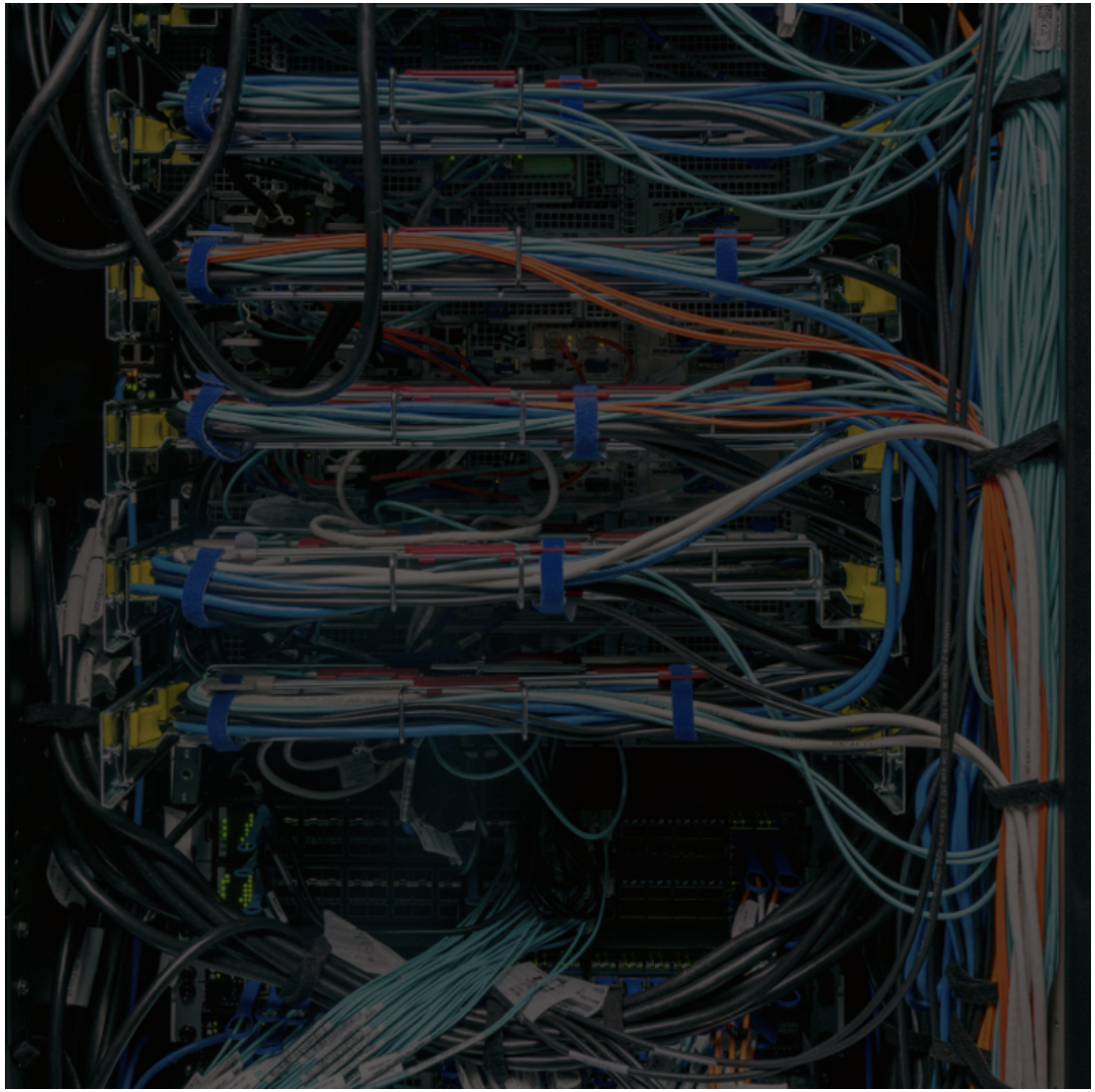
Ранг	**Система**	**ядер**	R_{peak}/R_{max} **(TFlop/сек)**	**Power (Мвт)**
1	Summit (Infiniband Melanox EDR) – USA	2414592	200794.9 / 148600.0	10.096
2	Sierra (Infiniband Melanox EDR) – USA	1572480	125712.0 / 94640.0	7.438
3	Sunway TaihuLight – China	10649600	125435.9 / 93014.6	15.371
4	Tianhe-2 (TH-IVB-FEP Cluster) – China	3120000	54902.4 / 33862.7	17.808
5	Frontera (Infiniband Melanox HDR) – United States	448448	38745.9 / 23516.4	
6	Piz Daint (Cray) – Switzerland	361760	25326.3 / 19590.0	2.384
7	Trinity (Cray) – USA	979072	41461.2 / 19135.8	7.578
8	ABCI (Infiniband EDR Fujitsu) – Japan	391680	32576.6 / 19880.0	1.649
9	SuperMUC-NG (Intel Omni-Path) – Germany	305856	26873.9 / 19476.6	
10	Lassen (Infiniband Melanox EDR) – USA	288288	23047.2 / 18200.0	
10* (2017)	K computer, SPARC64, Tofu interconnect – Japan	705024	11280.4 / 10510.0	12.660

Мощность Волховской ГЭС 86Мвт

Список ТОП-50 (Сентябрь 2019)

Ранг	**Система**	**процессоров/ядер**	R_{peak}/R_{max} **(TFlop/сек)**
1	МГУ (Ломоносов-2) – Т-платформы	1696/64384	4946.79 / 2478.0
2	Гидромецентр – Т-платформы, Cray	1952/35136	1293.0 / 1200.35
3	МГУ (Ломоносов-1) – Т-платформы	12422/82468	1700.21 / 901.90
4	Курчатовский интститут – Т-платформы	1070/21146	1100.55 / 755.53
5	Политех Санкт-петербург – РСК	1468/20552	1015.10 / 715.94
6	ВШЭ – Dell	68	912.4 / 568.5
7	Сколтех (ZHORES CDISE Cluster) – Dell	172	1011.6 / 495.9
8	АО "Тинькофф Банк"(Колмогоров)	20	658.5 / 418.9
9	МСЦ-РАН (МВС 10П) – РСК	416	523.83 / 383.21
10	ОИЯИ Дубна	496/9024	463.26 / 312.62

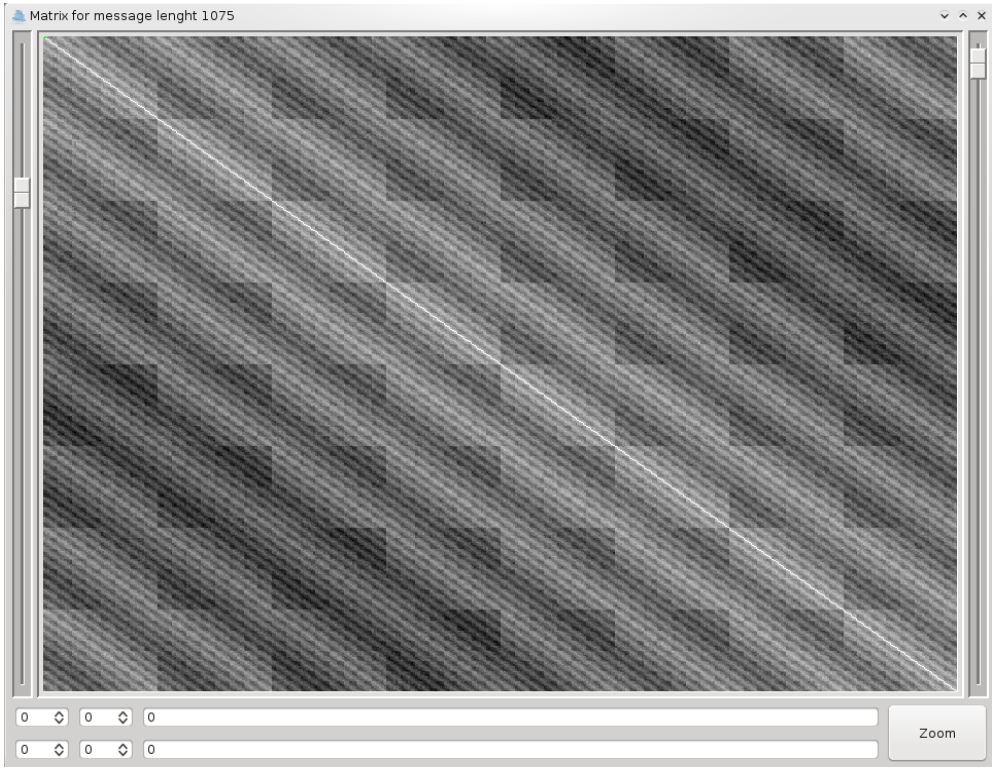
Типичная стойка в суперкомпьютере



Стойка в суперкомпьютере summit.

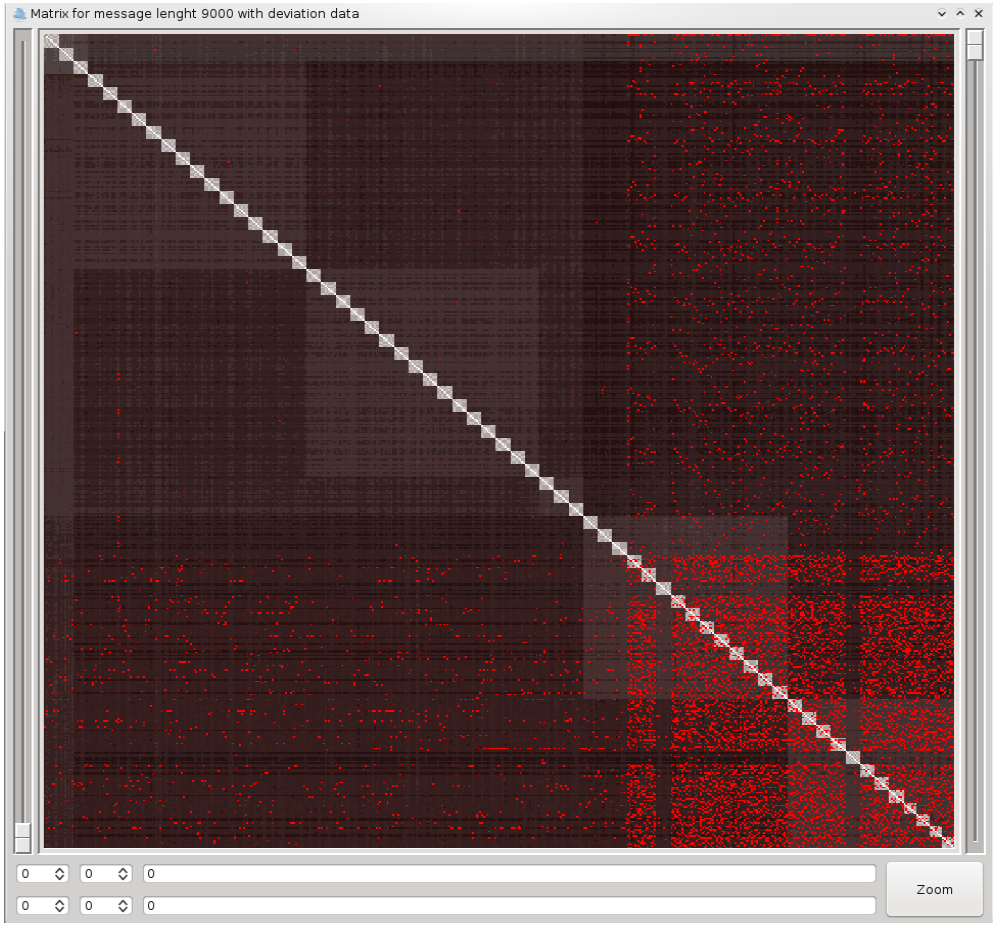
Топология коммуникаций в машине с BlueGene/P архитектурой

- “клеточная” структура определяется соседством процессоров в суперкомпьютере
- Линии параллельные главной диагонали определяются топологией трёхмерный тор.



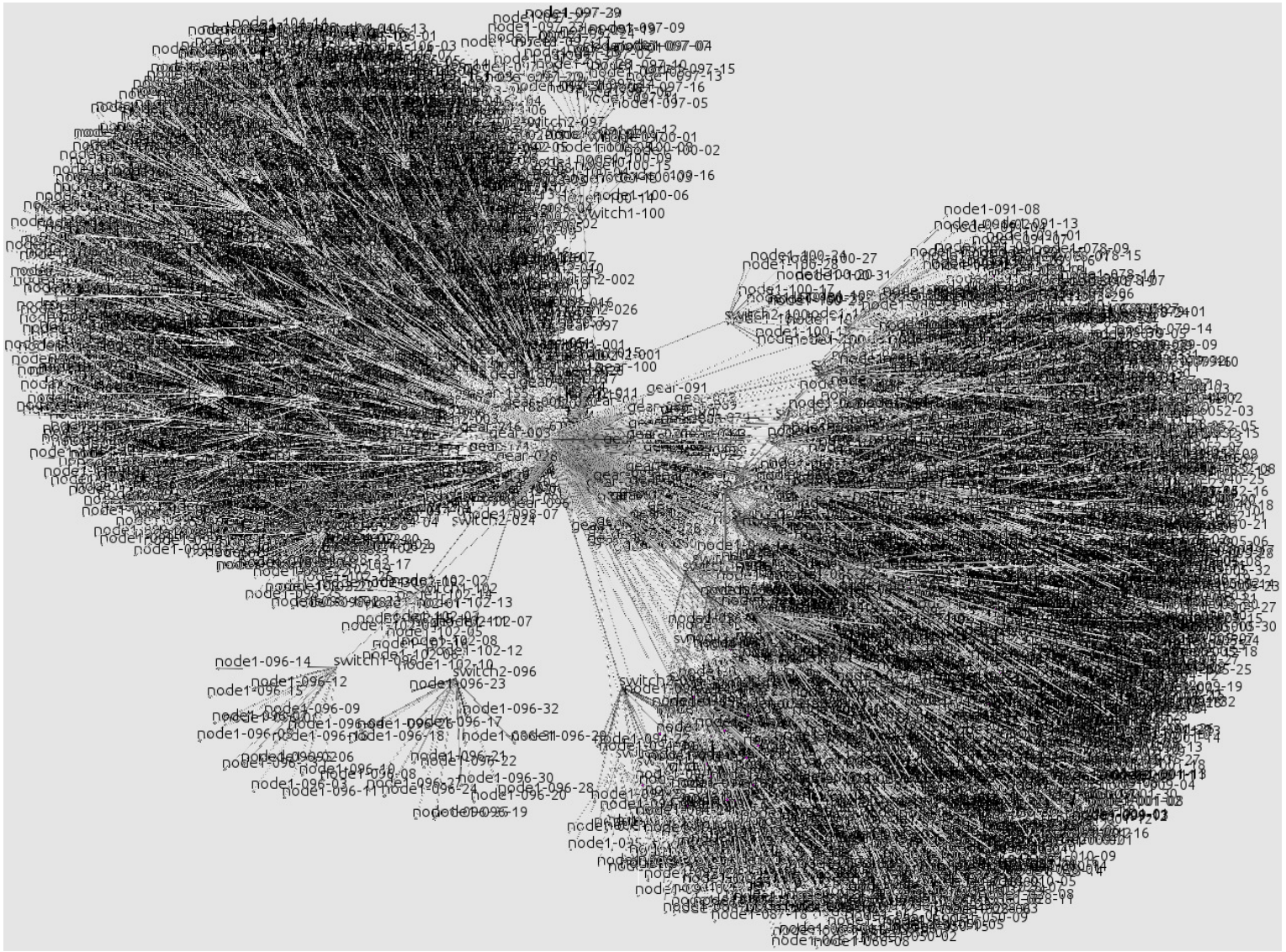
Суперкомпьютер “Ломоносов-1”

Распределение задержек при передаче для размера сообщения 9000 байт. Красное - степень вариативности в задержках.



Вариативность передач не равномерно распределена по суперкомпьютеру.

Структура коммуникационной среды “Ломоносов-1”



Система BlueGene/P

Для одних и тех же пар процессоров наблюдается различие во времени доставки сообщений. разница объясняется подключением топологии трёхмерный тор на 512 процессорах.

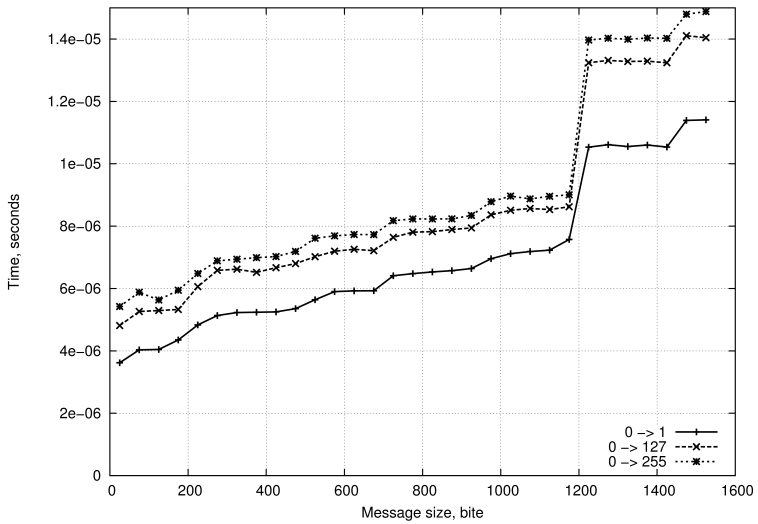


Рис.: 256 процессоров (1 процесс на 4 ядра).

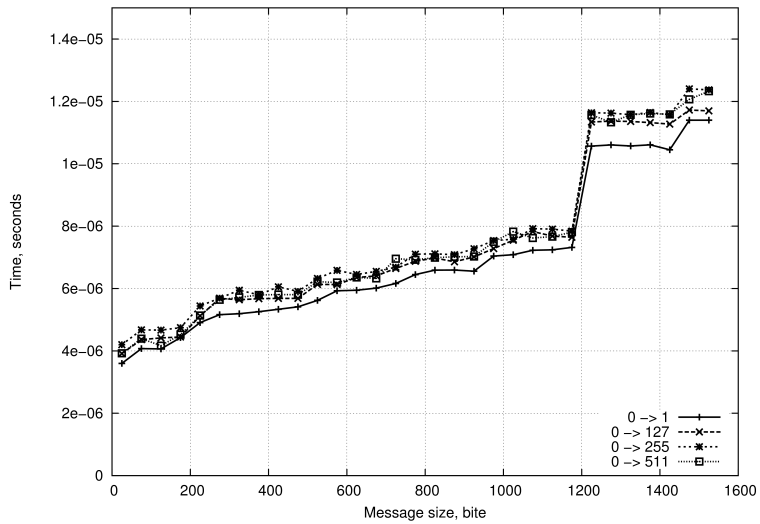
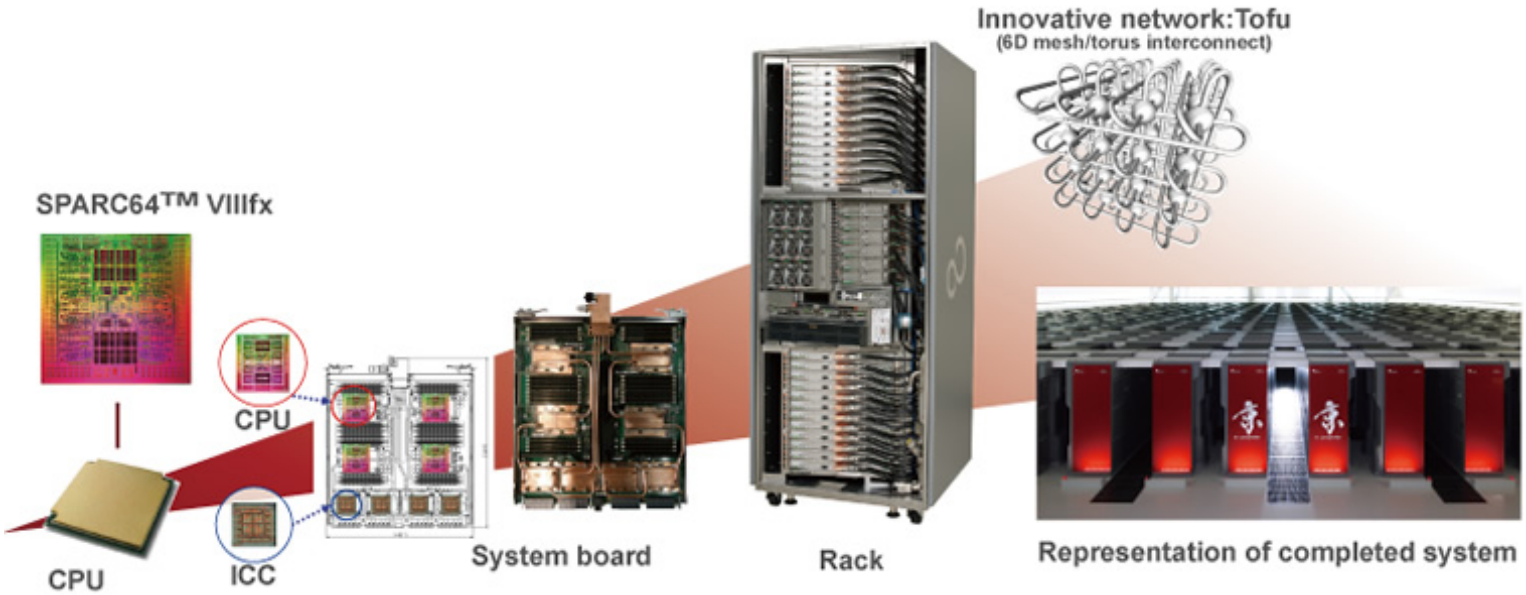


Рис.: 512 процессоров (1 процесс на 4 ядра).

Компоненты архитектуры K-компьютера



Что нужно знать про интерконнект в кластерах

- Сеть имеет регулярную структуру (топологию)
- Отказы в сети сравнительно редки (выход из строя узла наиболее частый отказ).
- Маршруты в сети обычно статические, в том числе обходы отказов.
- На самом деле сетей всегда несколько, под разные “коммуникационные паттерны” используются физически разные сети, в том числе сеть сервисная сеть, для управления и загрузки узлов кластера.

Всё это даёт возможность более-менее предсказывать задержку при передаче данных и решать задачу балансировки нагрузки в вычислениях.

О программном обеспечении вычислит. кластера

- Для программирования передач данных используется “протокол прикладного уровня MPI”.
- На узлах специализироваанный Linux, со специально доработанным ядром. (Для больших установок решается проблема сбоя узла, путём аппаратного включения запасного).
- Для пользователя система работает в пакетном режиме. Есть специальная система ведения очередей задач пользователей.

Направления деятельности

- **Сбор и анализ информации о коммуникационной среде**[1] Набор MPI-тестов, средства визуализации(приложение на QT+OpenGL), алгоритмы кластеризации. Языки описания коммуникационной среды.
- **Администраторская деятельность**. Создание систем для автоматизированного разворачивания сложной программной инфраструктуры, работа с системами ведения очередей и их анализ, модификация. (В том числе разворачивание прикладного ПО).
- **Распределённые вычисления**. Организация запуска потока задач, сразу на нескольких суперкомпьютерах.
- **Языки параллельного программирования**. Создание языка параллельного прогограммирования, в стиле программирование на графах.

[1] проект на github clustbench

Заключительный слайд с контактами

Спасибо за внимание!

- К работе над модификацией кода можно приступить здесь: `https://github.com/asalnikov`
- почта: salnikov@cs.msu.ru
- vk: asalnikov
- Сборник всякой формальной инвормации: `https://istina.msu.ru/profile/salnikov/`