Ростех Объединенная приборостроительная корпорация

Новые технологии высокоскоростной передачи данных

А.С. Семенов

План доклада

Сети

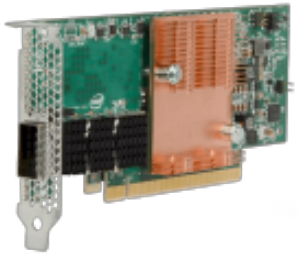
- Intel Omni Path
- Mellanox Infiniband EDR, HDR
- Tofu-2
- Atos BXI (Bull)

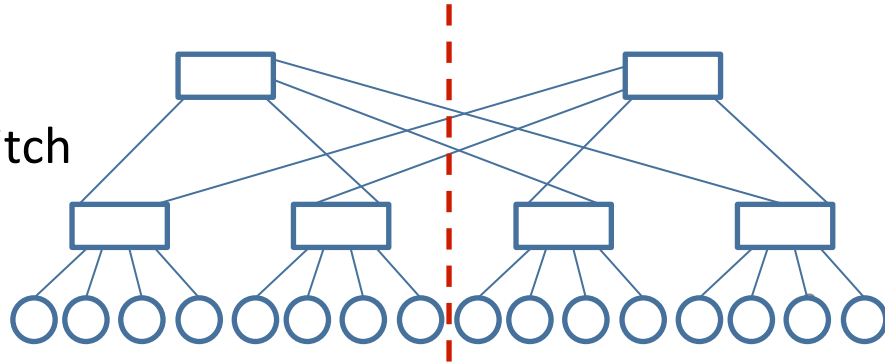
Аспекты

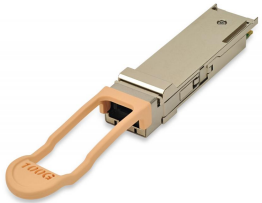
- Intel Omni Path vs Mellanox Infiniband
- Кластеры и суперкомпьютеры
- Жирное дерево vs тор

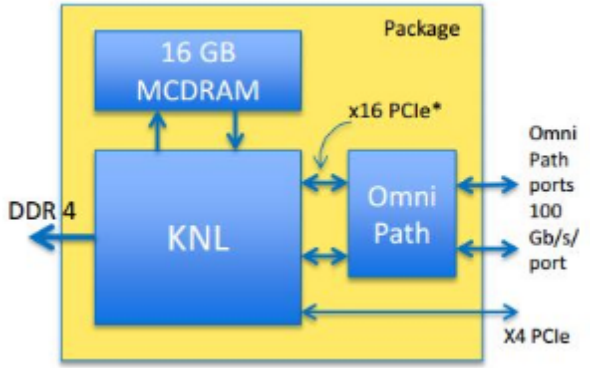
Intel Omni Path (2015)

- 100 Gbit/s,
- osu_latency 0.93 us (Haswell)
- 24/48 портов switch, 192/768 director-switch



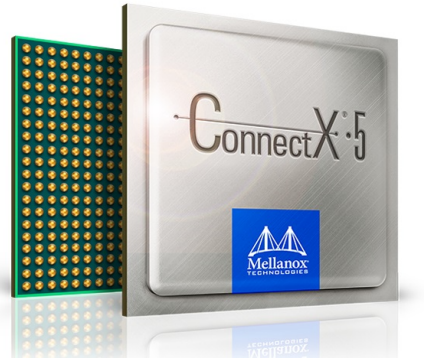






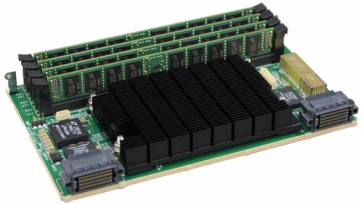
Infiniband EDR (2015)

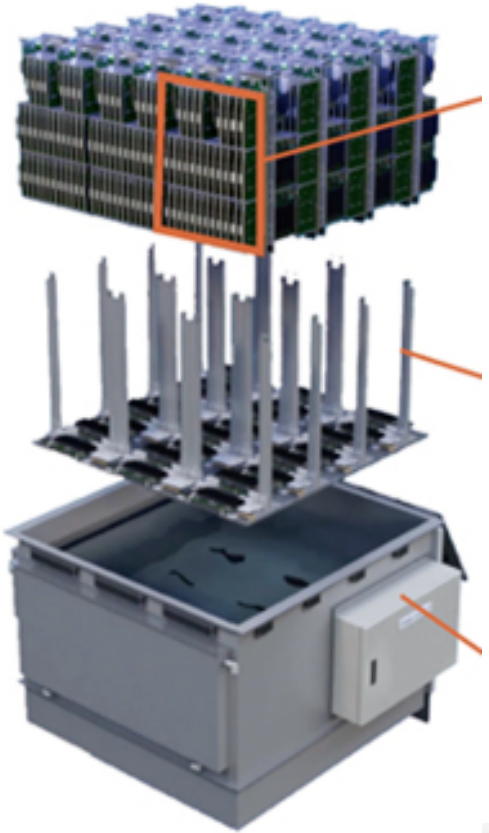
- 100 Gbit/s, osu_latency 1 us
- 36/216/324/648 портов



ZettaScaler-2.2 #4

- PEZY SC-2, 1 GHz, 1984-2048 C (8 way SMT), 4 TFLOPS, 180 W
- Intel Xeon D-1571 16C, 1.3 GHz (Broadwell)
- 10 000 PEZY SC-2 + 1250 x86 = 28.2 Пфлопс (67.9%)
- Mellanox ID EDR





Top500: Intel Omni Path vs Infiniband EDR

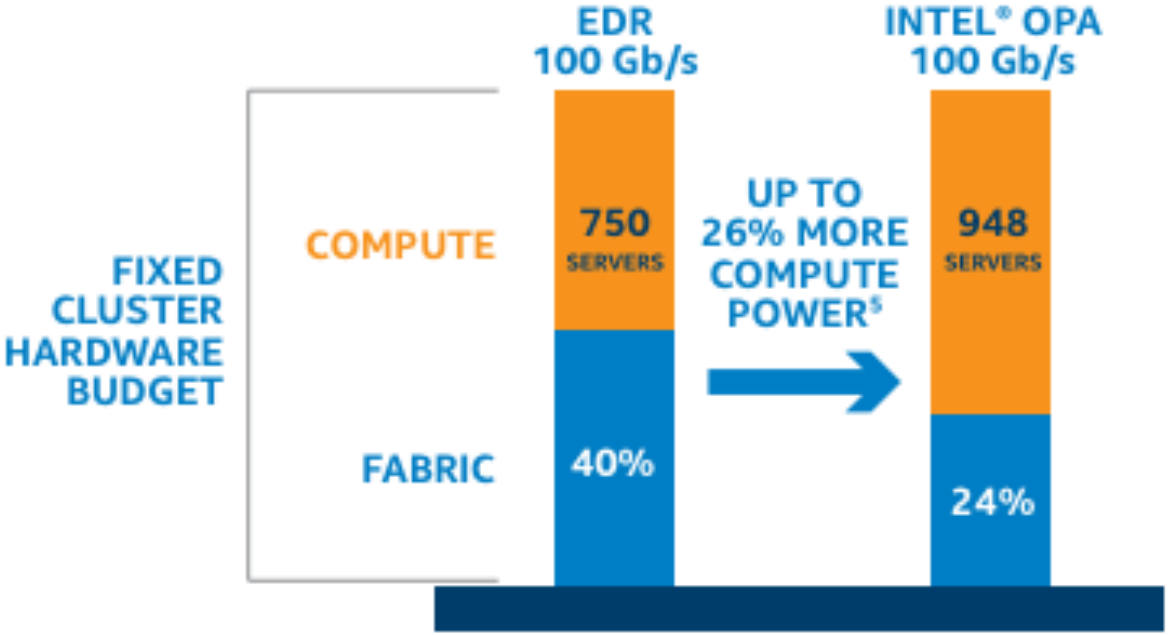
Intel Omni Path

- **#9** Intel Xeon Phi 7250, 1.4 GHz, 8178 узлов, 24.9 Пфлопс (54%, 1.2%)
- **#16** Barcelona SC MareNostrum4
  - Intel Xeon Platinum 8160 24C (Skylake, 2S), 2.1 GHz, 3192 узла, 10.3 Пфлопс (62.8%, 1.2%)
- в среднем по списку 61%
- 34 системы

Mellanox EDR

- **#4** ZettaScaler-2.2
  - 10 000 PEZY SC-2 + 1250 x86 = 28.2 Пфлопс (67.9%)
- **#17** Pleiades SGI ICE X
  - 7.1 Пфлопс (83.7%, 2.5%)
  - FDR
- 39 систем
- FDR – 105 систем

Intel Omni Path vs Infiniband EDR*



- https://www.intel.com/content/dam/www/public/us/en/documents/product-briefs/transforming-economics-hpc-fabrics-opa-brief.pdf

Intel Omni Path vs Infiniband EDR*

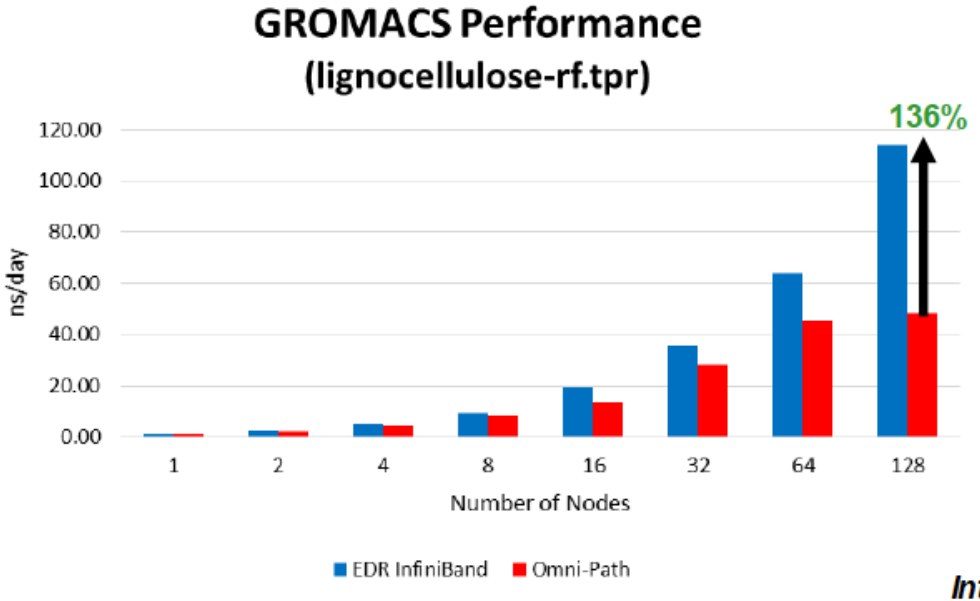
Operation	InfiniBand		Omni-Path	
	CPU Utilization	CPU Frequency at Operation Time	CPU Utilization	CPU Frequency at Operation Time
100Gb/s Data Throughput (Send-Receive)	0.8%	59%	59.6%	100%

Intel Performance Counter Monitor Tool - Output				
	Data Throughput (Gb/s)	AFREQ (relation to nominal CPU frequency while in active state)	CPU Instructions	Active Cycles
InfiniBand	99.5	0.59	39M	163M
Omni-Path	95	1	3725M	12000M

* - https://www.hpcwire.com/2016/06/18/offloading-vs-onloading-case-cpu-utilization

GROMACS Performance – Interconnects

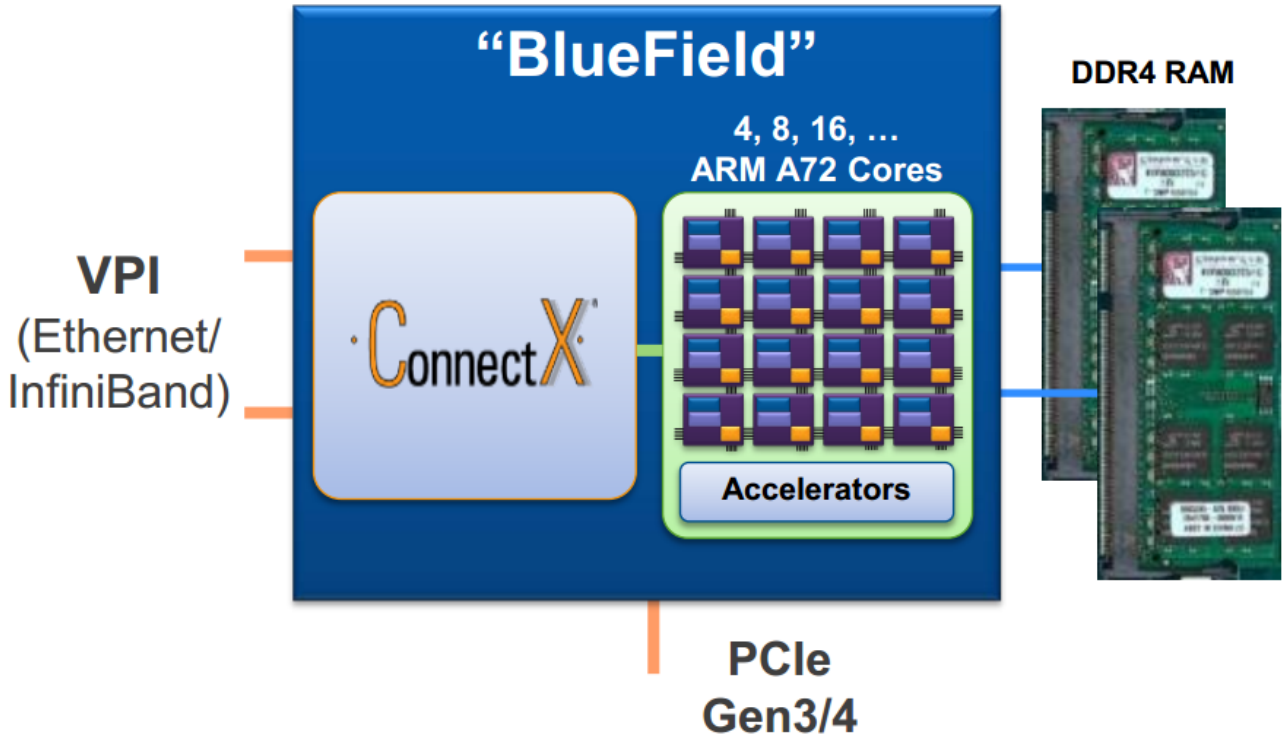
- **EDR InfiniBand enables higher scalability than Omni-Path for GROMACS**
  - InfiniBand delivers 136% better scaling versus Omni-Path for 128 nodes
  - 64 InfiniBand nodes delivers 33% higher performance compared to 128 Omni-Path nodes



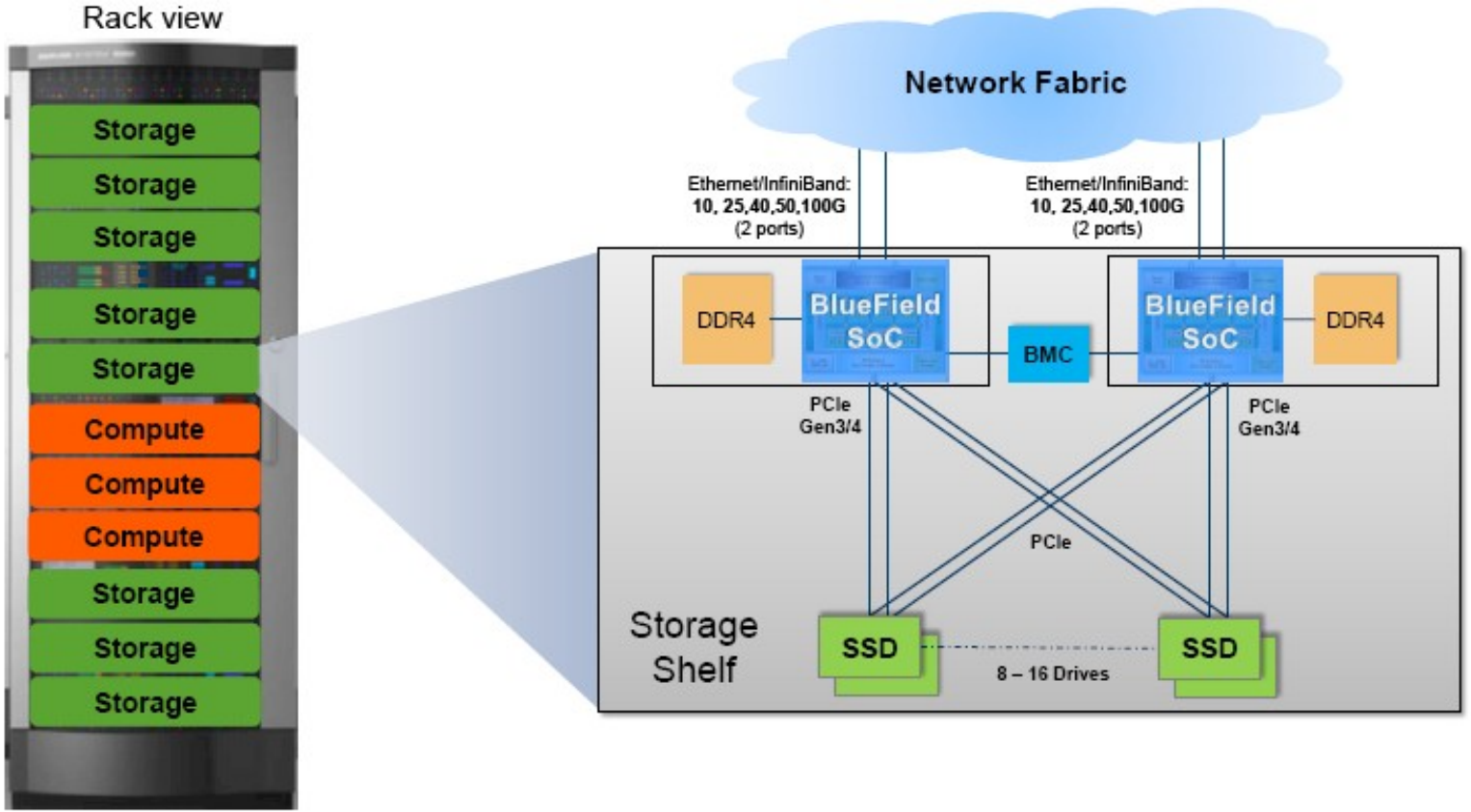
Higher is better

Intel MPI

Mellanox IB EDR Bluefield (100 Gbit/s)

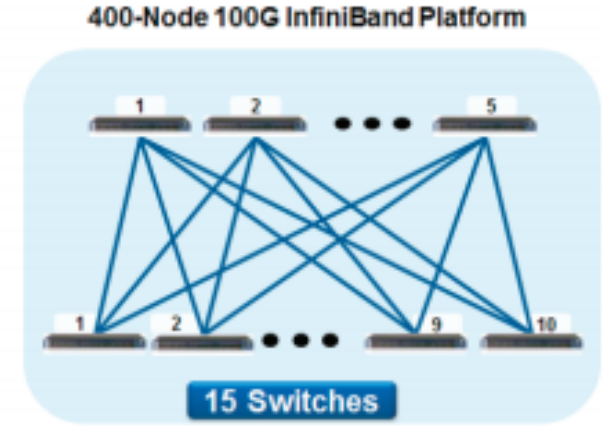


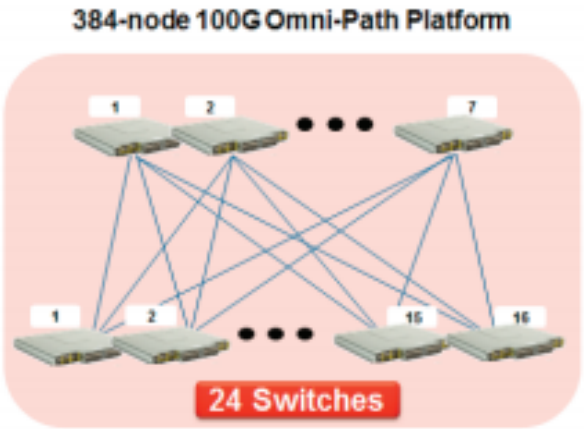
Mellanox IB EDR Bluefield (100 Gbit/s)



Infiniband HDR (2017)

- HDR 200 Gbit/s
- 40/200/800 портов



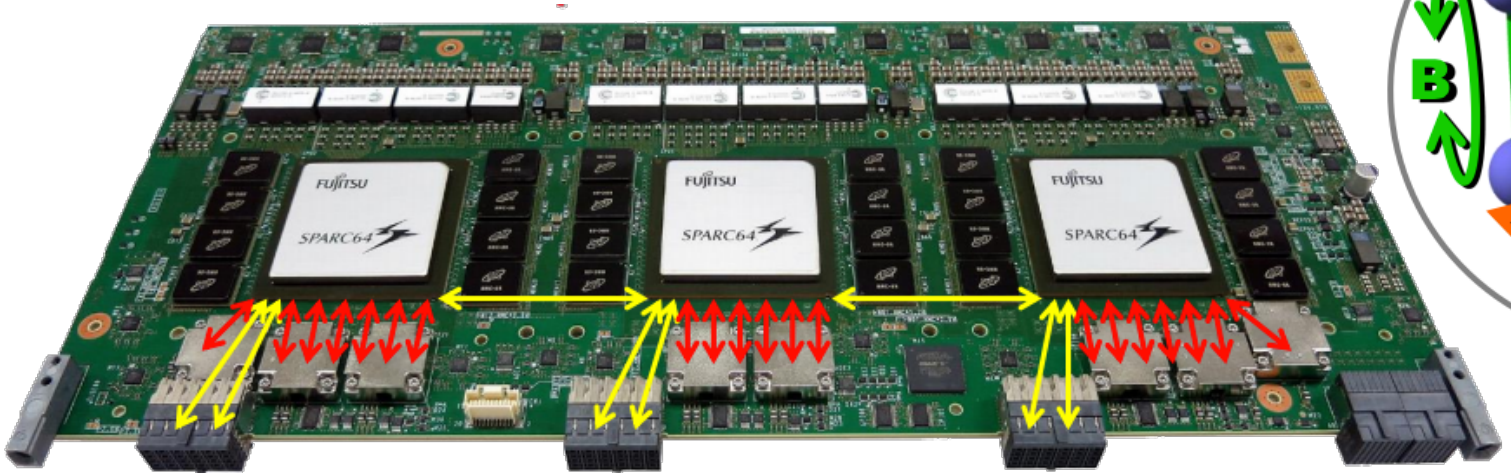


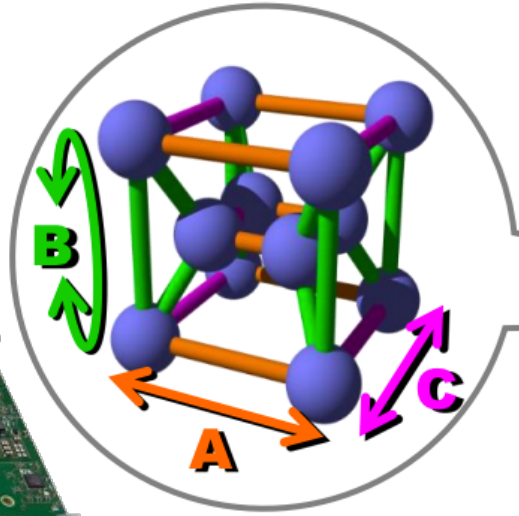
http://www.mellanox.com/related-docs/whitepapers/WP_Introducing_200G_HDR_InfiniBand_Solutions.pdf

Tofu-2 Interconnect (2015)

- Fujitsu, PRIMEHPC FX100
- 6D-тор, 100 Gbit/s
- 4 интерфейса, 10 линков
- Задержка 0.71 мкс (low level? MPI?)
- 6/7 **оптических линий**, 3/4 электрических

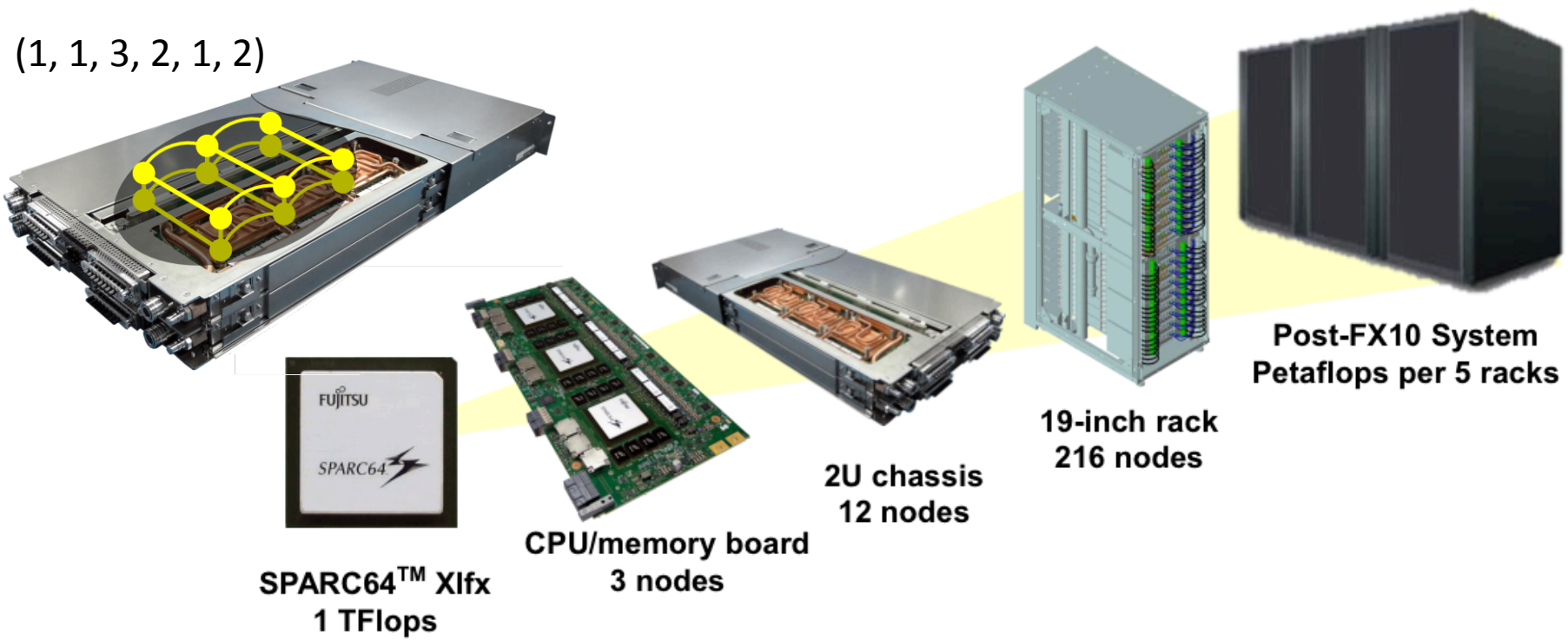
- SPARC64 XIfx, SoC, 2.2 GHz
  - 2 FMA, 256 bits, 1.1 Тфлопс,
  - 8 HMC (120 GB/s), 20 nm



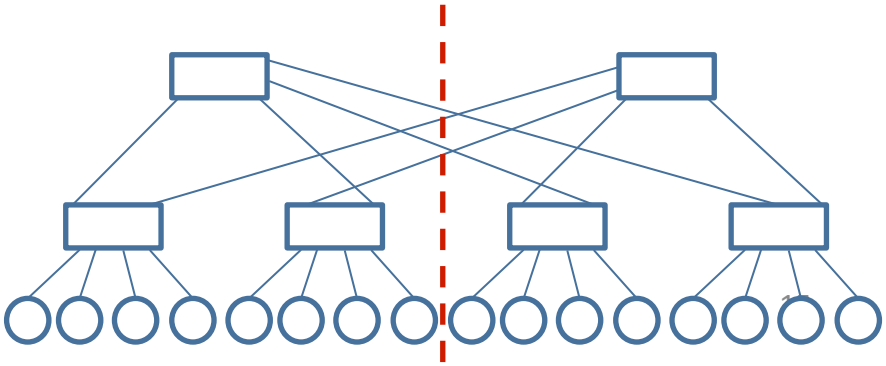


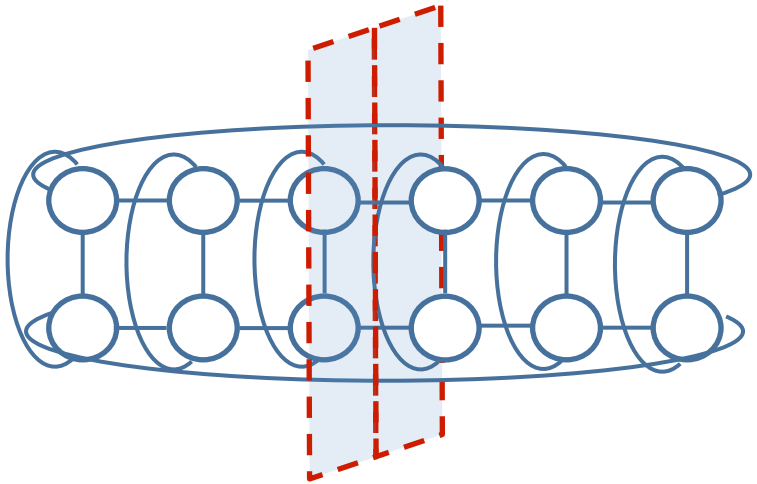
Tofu-2 Interconnect (2015)

(1, 1, 3, 2, 1, 2)



- **#38** Nov 2017 Top500: SPARC64 XIfx 32C, 3.5 Пфлопс (90.7%, 3.2%)
- 5 систем





Бисекция жирного дерева (half bisection) = N/4

Бисекция тора = 2N/Nmax

Bull Exascale Interconnect (BXI)

- 100 Gbit/s, PCIe gen 3 x16
- до 64 000 узлов
- Несколько виртуальных сетей, Апаптивность
- 2 ASIC − switch, NIC
- BXI доступно как мезанин − для Bull Sequana, как PCIe карта
- Маршрутизация: каждый порт имеет свою таблицу маршрутизации. Для каждого узла назначения можно определить 3 выходных порта: 1 детерминированный, 2 адаптивных
- Глобальный механизм синхронизации
- Portals4 − non-connected протокол, минимизация использования памяти
- Коммутаторы
  - 48 портов (тоже с жидкостным охлаждением на горячей воде), медные кабели
  - 288 или 576 портов − 2 уровня ASIC

Bull atos technologies

Bull Exascale Interconnect (BXI)

- Bull Sequana − платформа с жидкостным охлажде
  - в стойке до 288 узлов, 2 уровня топологии Fa
  - возможна установка 2 портов на узел
  - KNL

- **#23** Nov 2017 Top500: 7250, 9.3 Пфлопс, 3072 узла
- Всего 17 систем

Сравнение основных характеристик

Сеть	Mellanox IB FDR 4x	Ангара	Mellanox IB EDR 4x	Intel OmniPath	Mellanox IB HDR 4x	Ангара-2
Год выпуска	2011	2013	2015	2015	2017	2018
№ в Top500	17	–	4	9	–	
Топология	fat tree / kD-тор	4D-тор	fat tree / kD-тор	fat tree	fat tree	модиф. 4D-тор
Задержка MPI, мкс	1	0.85	1	0.93	н/д	< 0.8
Задержка на хоп сети, нс	– / 250	129	н/д	н/д	н/д	< 100

Special Issue of Lobachevskii Journal of Mathematics on "High Performance Data Intensive Computing"

НИЦ ИВТ

Ростех Объединенная приборостроительная корпорация

Topics of Interest

- Data-intensive applications and High Performance Computing, Big Data, Artificial Intelligence
- Hardware architecture for data-intensive computing in exascale era
- Power and energy efficiency, storage and file systems with regard to data-intensive applications
- Specific issues of performance evaluation, analytic modeling, simulation for data-intensive applications

Scopus®

Web of Science®

Important Dates

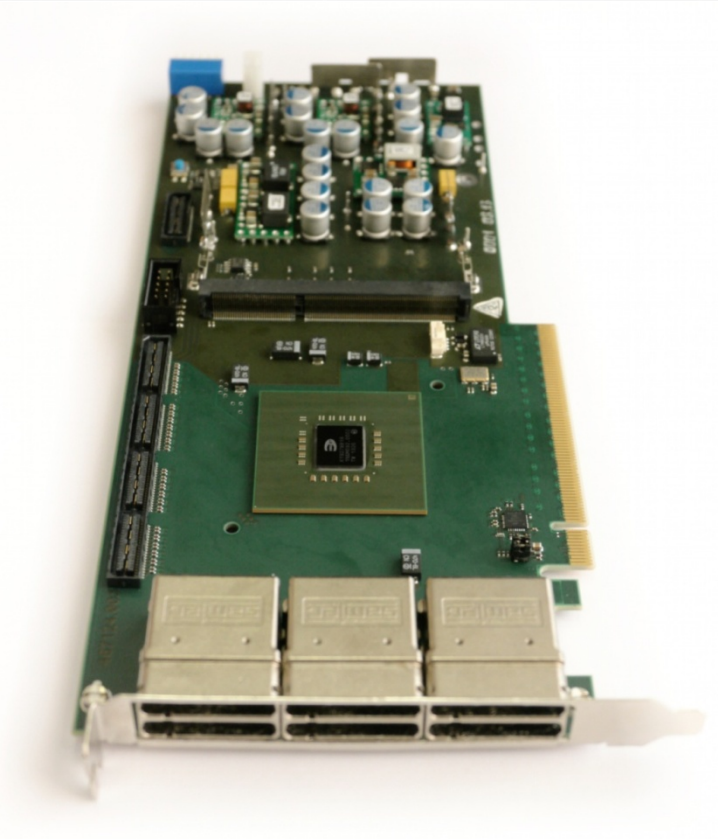
- **December 15th, 2017** title and abstract submission
- **January 22th, 2018** paper submission
- 1st general review round
- **February 19th** author notification
- **March 12th** camera ready submission edited after reviewers comments

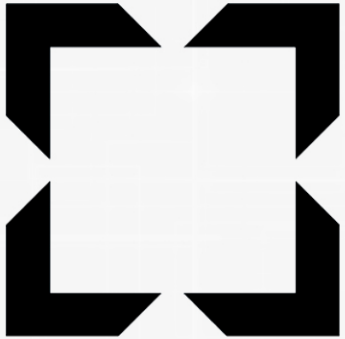
http://dislab.org/ljm_cfp2018

Контакты:

117587, Москва, Варшавское ш, 125

angara@nicevt.ru





Ростех

Объединенная
приборостроительная
корпорация