СУПЕР-ЭВМ

СЕТЕВЫЕ РЕШЕНИЯ

Комаров С.О.

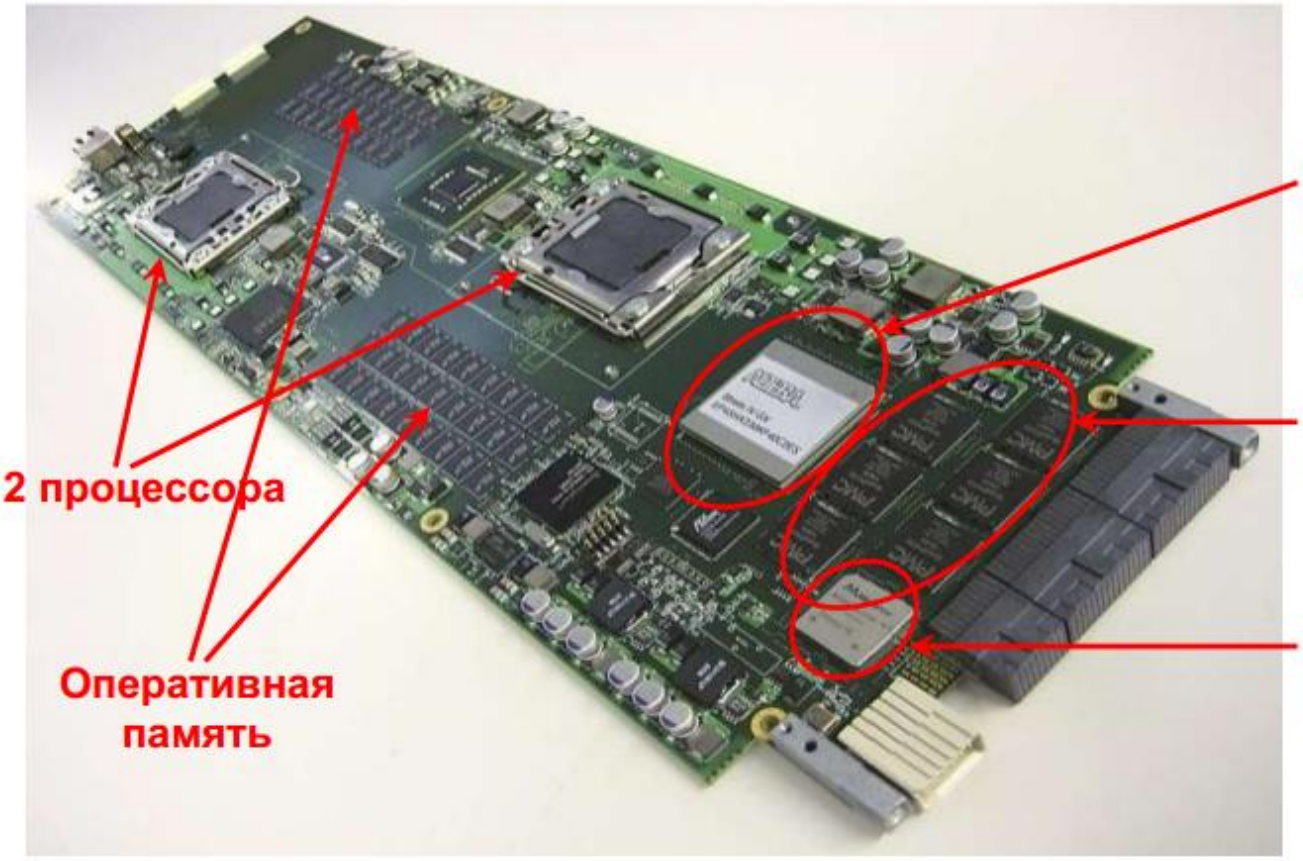
КОММУНИКАЦИОННЫЕ СЕТИ СУПЕРКОМПЬЮТЕРОВ

- Коммерческие:
  Infiniband
- Не коммерческие:
  IBM BlueGene, Cray XT
- Российские разработки:
  СКИФ-Аврора, МВС-Экспресс, Ангара

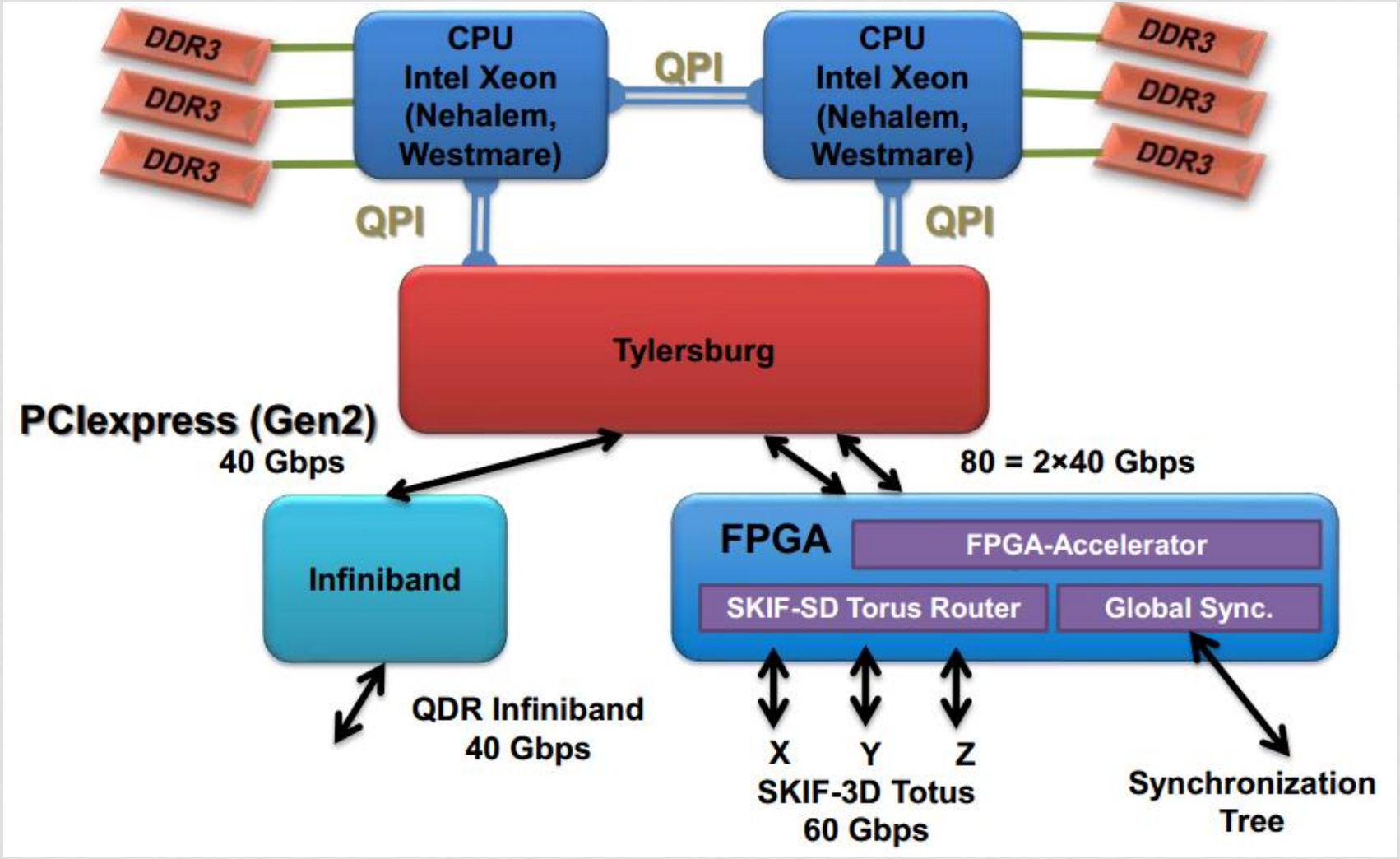
СЕМЕЙСТВО “СКИФ”

Ряд	Годы и пиковая произв-ть	ядер CPU/ разряд.	Сетевые решения	Форм-фактор **CPU/U**	Примечание
1	2000–2003 20–500 GFlops	1/32	FastEth, SCI (2D-тор), Myrinet	4U–1U **0.5–2**	Отечественный SCI (2D-тор)
2	2003–2007 0.1 – 5 Tflops	1/ 32–64	GB Eth, SCI (3D-тор), InfiniBand	1U, HyperBld. **2**	ServNet v.1, v.2 Ускорители: FPGA, ОВС
3	2007–2008 5–150 Tflops	2–4/64	GB Eth InfiniBand DDR	1U, blades **2–4**	ServNet v.3 воздух—вода—фреон
4	2009–2012 ~0.5–5 Pflops	4–12/64	InfiniBand QDR, **отечественная сеть (3D-тор)**	**плотные** blades **10.7**	Ускорители: FPGA, GPU, МЦОС

СКИФ-АВРОРА

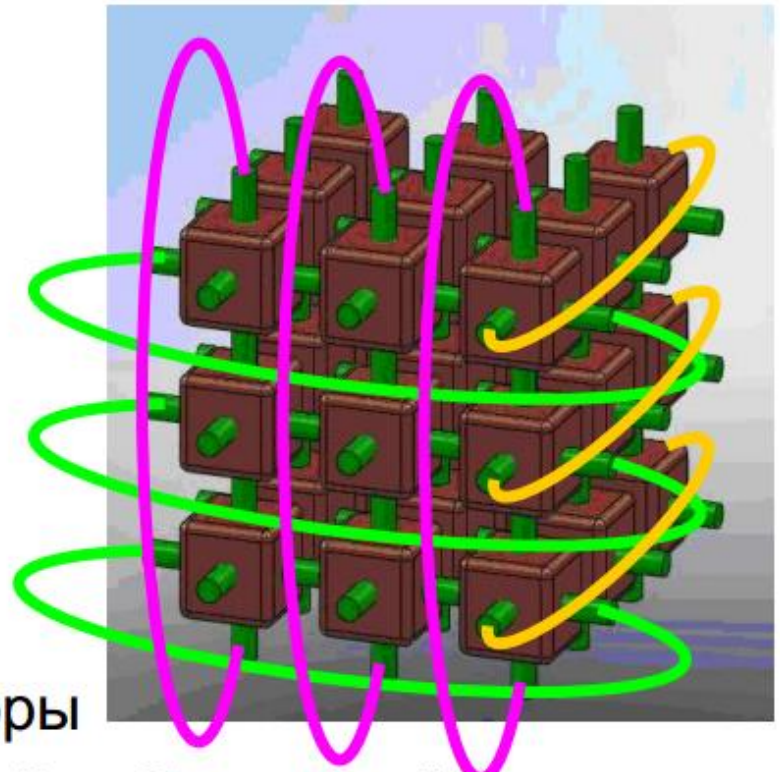

2 процессора
Оперативная память
ПЛИС маршрутизатора системный сети «трехмерный тор»
6 трансиверов системной сети «трехмерный тор»
Вспомогательная сеть InfiniBand

СКИФ-АВРОРА


DDR3
DDR3
DDR3
CPU
Intel Xeon
(Nehalem,
Westmare)
QPI
CPU
Intel Xeon
(Nehalem,
Westmare)
DDR3
DDR3
DDR3
QPI
QPI
Tylersburg
PCIexpress (Gen2)
40 Gbps
80 = 2×40 Gbps
Infiniband
FPGA
FPGA-Accelerator
SKIF-SD Torus Router
Global Sync.
QDR Infiniband
40 Gbps
X
Y
Z
SKIF-3D Totus
60 Gbps
Synchronization
Tree

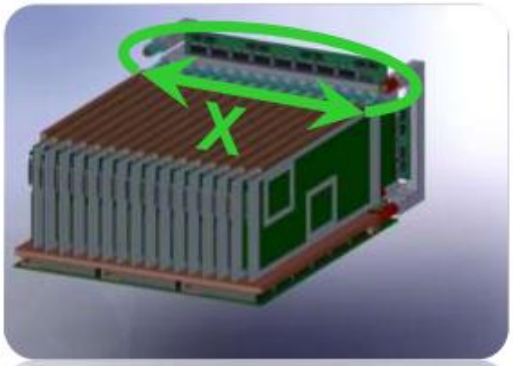
СИСТЕМНАЯ СЕТЬ «3D-TOP»

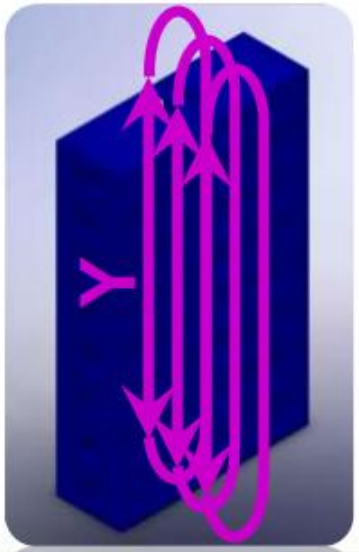
- ☆ Хорошие: пропускная способность, задержка и темп выдачи сообщений
- ☆ Перестраиваемая системная сеть
  - Коммутаторы для конфигурирования сети
  - Тор можно разбить на подторы
  - Задаче можно назначить свой собственный подтор

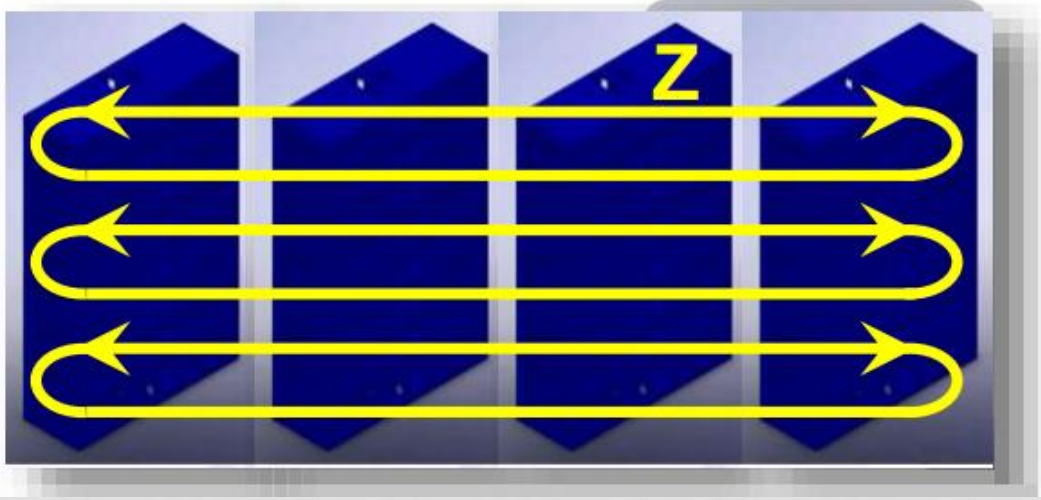


СИСТЕМНАЯ СЕТЬ «3D-TOP»

- 1ое направление (X): 16 узлов, соединение без кабелей, на соединительной панели
- 2ое направление (Y): 16 шасси, кабели внутри стойки
- 3ье направление (Z): кабели между стойками, 1–32 стойки







ПРОГРАММНЫЙ СТЕК SKIF-3D-TORUS

SKIF-3D-ROUTER (VHDL)

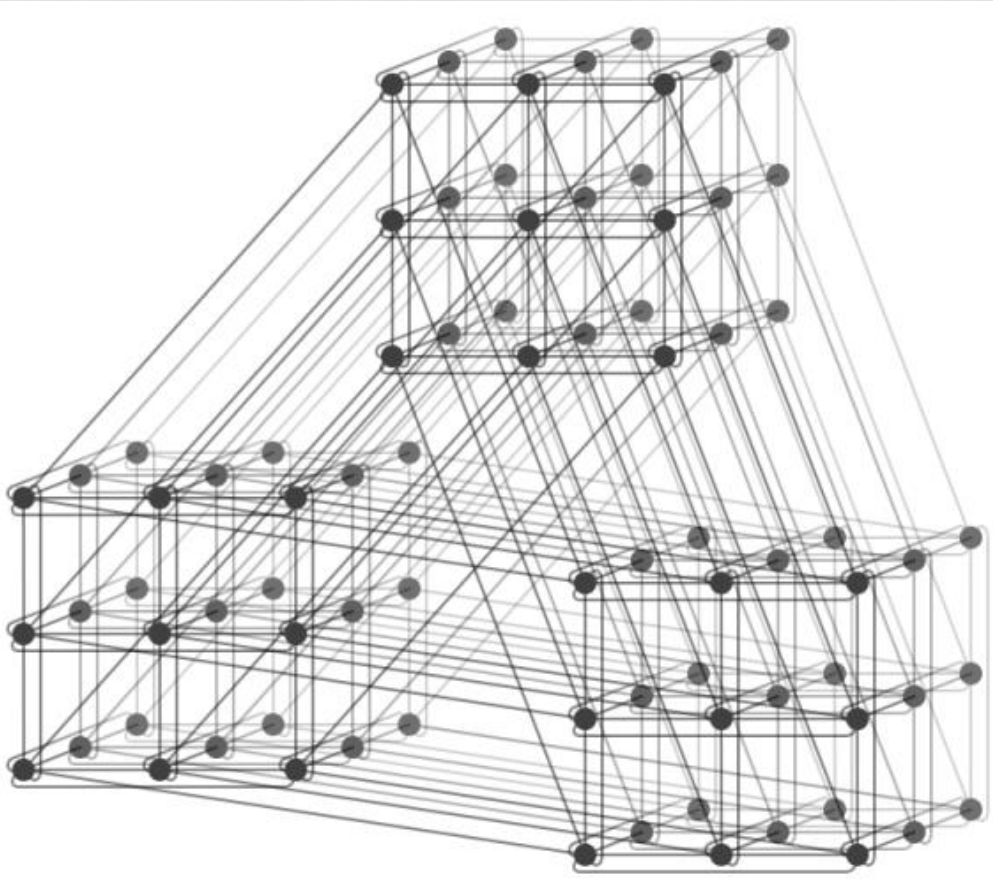
- Роутер реализован в FPGA
  - решены проблемы вида “deadlock”, “livelock”, “starvation”
  - виртуальные каналы и туннели
  - отказоустойчивость
- Реализованы различные алгоритмы маршрутизации и их композиции
  - Статическая маршрутизация
  - Адаптивная маршрутизация
- “SkifCh”— коммуникационная библиотека низкого уровня

TESTING

Убраны громоздкие и неэффективные схемы трансляции адресов, присутствующие в коммерческих коммуникационных сетях, таких как Infiniband.

	SKIF-3D-Torus	Infiniband QDR
Bandwidth (Gbps per node)	60	40
Message Rate (MT/s)	14	3
Latency (μs)	1–1.5	1–1.5

КОММУНИКАЦИОННАЯ СЕТЬ “АНГАРА”



- Топология «4D-тор»
- Односторонние коммуникации:
  - put (запись в удалённую память)
  - get (чтение из удалённой памяти)
  - атомарные операции add и xor
- Коллективные операции:
  - broadcast
  - reduce
- 10 инжекционных конвейеров
- Адаптивная передача пакетов
- Механизмы синхронизации

КОММУНИКАЦИОННАЯ СЕТЬ “АНГАРА”

Характеристика	Ангара (FPGA)	Ангара (ASIC)	Infiniband FDR 4x	IBM BG/Q (Custom)	Cray XK7 (Custom)
Топология сети	**2D-тор**	**4D-тор**	Fat tree	5D-тор	3D-тор
ПС с процессором, Гбайт/с	**2**	**8**	6,4	~ 20	9,6
ПС линка, Гбайт/с	**0,78**	**7,5**	6,8	2	9,375
Агрегатная ПС линков, Гбайт/с	**6,3**	**120**	-	40	186
Задержка между соседними узлами, мкс	**2,1**	**1.0**	1,0	< 1,0	1,4
Стоимость, т. руб./узел	**≈120**	**≈38**	≈41	-	-

КОММУНИКАЦИОННАЯ СЕТЬ “АНГАРА”

Характеристики ЕС8430:

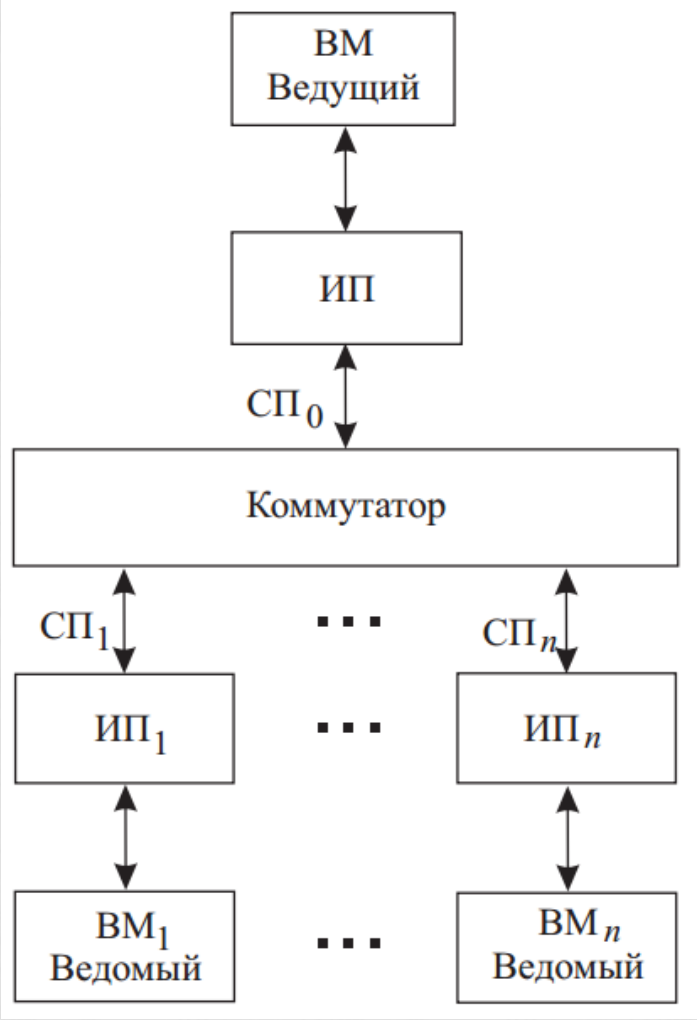
Техпроцесс…….…… TSMC 65nm GP
Размер……13.0mm x 10.5mm
Кол-во транзисторов……..……180M
Частота……..……..……500MHz
TDP……..……36W
Интерфейсы:
GEN II PCI-E…..……x16
(5.0Gbps/lane, 80 Gbps total each way)
Links……..……x8 (1-12 lanes/link
3.125-6.25 Gbps/lane, max. 75 Gbps/link
each way, total max. 600 Gbps each way)
DDR3 SDRAM……..Peak BW 8.5 Gbyte/s
(72 bit, 1066 MT/s)
Электропитание:
SerDes ……..1.0V±5%
Core……..……..1.0V±5%
I/O……..…2.5V±10%
Темп. диапазон……..……0-70°C
Корпус…..……FCBGA-1521 40mm x 40mm

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ “АНГАРА”

- Поддержка прикладных (математических, алгоритмических) библиотек (BLAS, LAPACK, FFTW, GSL, MKL и т.д.).
- Поддержка компиляторов языков Fortran 77/90/95 (GNU, Intel), C/C++ (GNU, Intel), UPC, Co-Array Fortran.

КОММУНИКАЦИОННАЯ СЕТЬ НА БАЗЕ «МВС-ЭКСПРЕСС»

- Коммутатор
- Интерфейсные платы (ИП), подключенные к магистралям PCI Express вычислительных модулей (ВМ)
- Сетевых кабелей (СП)

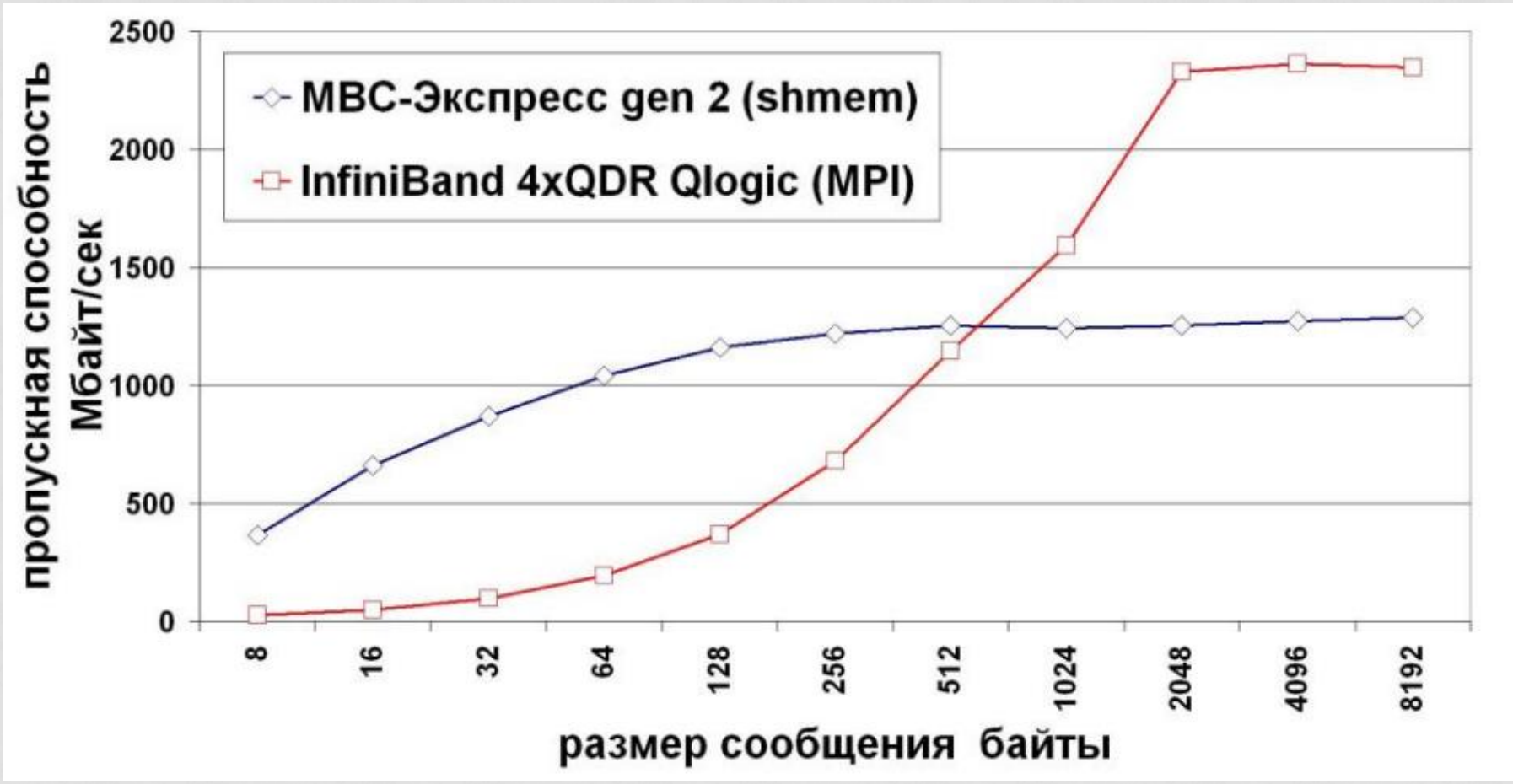


ОСОБЕННОСТИ «МВС-ЭКСПРЕСС»

- Использование PCI Express для передачи данных как внутри узла, так и между
  - должен быть выделен один ВМ, называемый «ведущим», в отличие от остальных - «ведомых», для размещения в его адресном пространстве окон доступа ко всем ведомым узлам
  - в каждом ВМ, как правило, должен быть выделен тред или ядро (в многоядерных ВМ), предназначенные для работы с интерфейсной платой, для поддержки синхронизации и когерентного состояния памяти ВМ.

ПРОГРАММНЫЙ СТЕК «МВС-ЭКСПРЕСС»

TESTING


МВС-Экспресс gen 2 (shmem)
InfiniBand 4xQDR Qlogic (MPI)
пропускная способность Мбайт/сек
2500
2000
1500
1000
500
0
8
16
32
64
128
256
512
1024
2048
4096
8192
размер сообщения байты

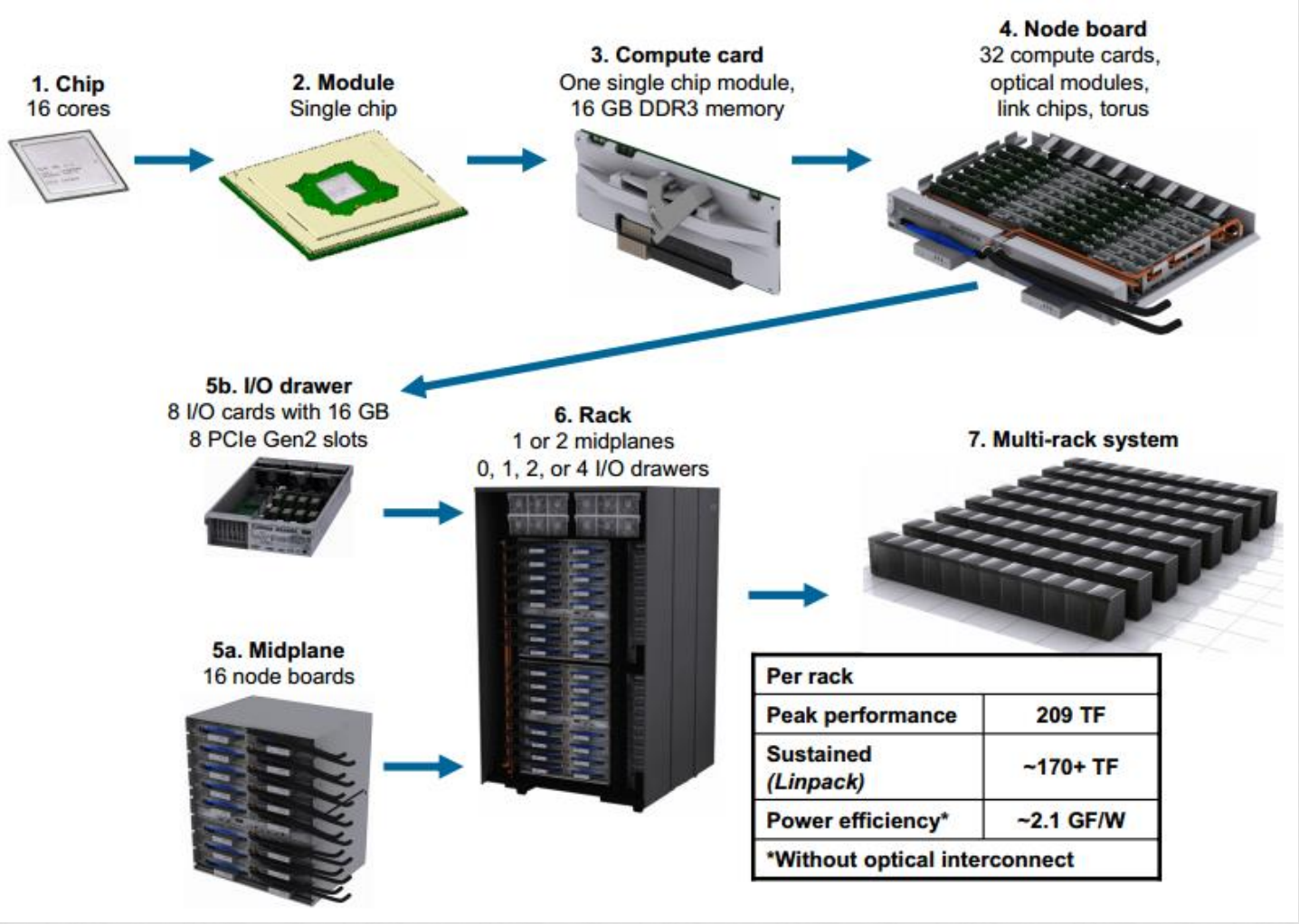
BLUE GENE

- IBM Blue Gene Projects:
  - **Blue Gene / L (280 TFLOPS)**
  - **Blue Gene / C**
  - **Blue Gene / P (1 PFLOP)**
  - **Blue Gene / Q (3-10 PFLOPS)**

BLUE GENE / P

- 3D Top
  - сеть общего назначения, объединяющие все вычислительные узлы (p2p)
  - пропускная способность — 425 MB/s
  - латентность (ближайший сосед):
    - 32-байтный пакет: 0,1 μs
    - 256-байтный пакет: 0,8 μs
- глобальные коллективные данные
  - коммуникации типа «один-ко-многим» (broadcast/редукция)
  - пропускная способность — 850 MB/s
  - латентность (полный обход): 3,0 μs
  - операции барьеров и прерываний (глобальные AND- и OR-операции)

BLUE GENE / Q



Per rack	
Peak performance	209 TF
Sustained *(Linpack)*	~170+ TF
Power efficiency*	~2.1 GF/W
*Without optical interconnect	

BLUE GENE / Q

- 5D Top — 40 GB/s
- Латентность 2.5 мкс
- Broadcast/Reduce – часть 5D Top
- PCIe x8 Gen2 based I/O
- 1 GB Control Network — System Boot, Debug, Monitoring