TECHNICAL SUPERCOMPUTERS | LABORATORY

TECHNICAL SERVERS | LABORATORY

TECHNICAL STORAGE | LABORATORY

Построение сбалансированного кластера

ООО «Терминал-Сервис»,
Сергей Дудинов,
Директор
sd@tslab.com.ua

Предположения (реалии Украины)

- Мы не строим суперкомпьютер в общемировом понятии - Бюджет максимально ограничен
- Есть инфраструктура – место для установки, электропитание, человеческие ресурсы для обслуживания
- Цель – получить максимум терафлопс на гривну сейчас, предусмотреть возможности простой модернизации

Выбор железа – что лучше?

- CPU или GPU?
- Платформа
- Процессоры – 1, 2, 4 или 8 на ноду?
- Интерконнект – Ethernet или Infiniband?
- Много слабых нод или мало, но мощных?

CPU или GPU?


intel
Xeon E5-2600

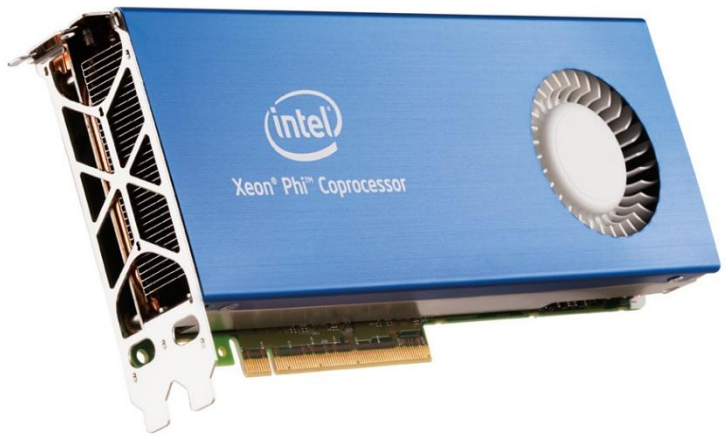

NVIDIA TESLA


AMD
64
Opteron


intel

NVIDIA


AMD


intel
Xeon Phi Coprocessor

RADEON
GRAPHICS
S10000

CPU или GPU?

Каждая из этих двух архитектур имеет свои достоинства. CPU лучше работает с последовательными задачами.

При большом объеме обрабатываемой информации очевидное преимущество имеет GPU.

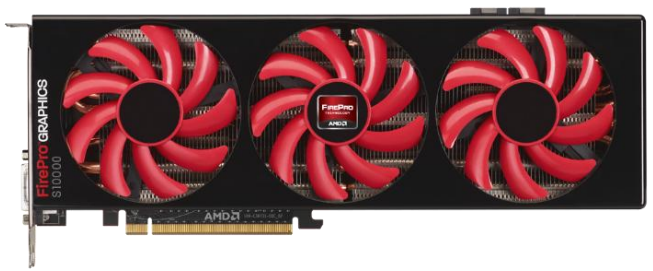
Условие только одно — в задаче должен наблюдаться параллелизм.

CPU или GPU?

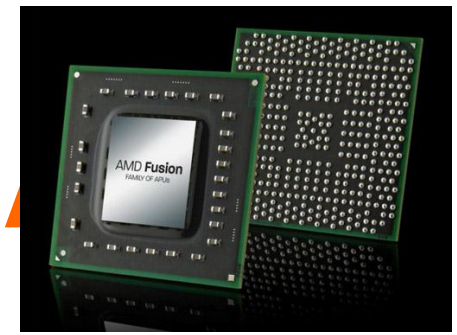
GPU уже достигли той точки развития, когда многие приложения реального мира могут с легкостью выполняться на них, причем в разы быстрее, чем на самых топовых CPU. Будущие вычислительные архитектуры станут гибридными системами с графическими процессорами, состоящими из параллельных ядер и работающими в связке с многоядерными ЦП

CPU или GPU?

Подтверждение тому – проекты «AMD Fusion» и «VIA CoreFusion». Суть – объединение центрального многозадачного универсального процессора с графическим параллельным многоядерным процессором в одном кристалле.



=



CPU или GPU?

HPC процессоры будущего будут строиться на гибридной архитектуре

Вычисления на GPU ускорителях:

- Становятся универсальнее и удобнее для программирования;
- Находятся в технологическом авангарде с поддержкой со стороны массового рынка;
- Признаны всеми игроками рынка как правильное направление развития.

Вывод: кластер собираем только с GPU !!!

Какой GPU?

Ключевые возможности	Архитектура GPGPU	Пиковая производительность, Гигафлоп		Полоса пропускания памяти (без ECC), ГБ/с	Размер памяти (GDDR5), ГБ	Ядра, шт
		для вычислений двойной точности с плавающей точкой	для вычислений одинарной точности с плавающей точкой			
Xeon Phi Coprocessor 5110P	1 Many Integrated Core	1010	н.д.	320	**8**	60/240
TESLA M2075	1 Fermi GPU	515	1030	150	6	448
TESLA M2090	1 Fermi GPU	665	1331	177	6	512
TESLA K10	2 Kepler GK104s	190 (95 на 1)	4577 (2288 на 1)	320 (160 на 1)	**8** (4 на 1)	3072 (1536 на 1)
TESLA K20	1 Kepler GK110	1170	3520	208	5	2496
TESLA K20X	1 Kepler GK110	1310	3950	250	6	2688
FireStream 9370	1 RV870	528	2640	153	4	1600
FirePro S9000	1 Tahiti LE	806	3230	264	6	1792
FirePro S10000	2 Tahiti LE	**1480** (740 на 1)	**5910** (2955 на 1)	**480** (240 на 1)	6 (3 на 1)	**3584** (1792 на 1)
FirePro W9000	1 Tahiti XT	998	3993	264	6	2048

Какой GPU?

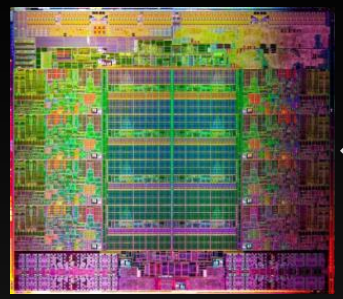
Accelerator/Co-Processor	Rmax (GFlops)*	Rpeak (GFlops)	Эффективность, RMax/Rpeak	Реальная производительность 1 ускорителя, Гигафлоп	
				Double prec.	Single prec.
TESLA M2090	4399354	10002202	44,0	292,6	585,6
TESLA M2050	7204750	14143446	50,9	262,1	524,3
TESLA M2070	1529010	2777116	55,1	283,8	567,5
TESLA K20x	17863700	27505427	64,9	**850,2**	2563,6
TESLA K20	н.д.	н.д.	**71,0***	830,7	2499,2
TESLA K10	н.д.	н.д.	**71,0***	134,9	**3249,7**
Xeon Phi 5110P	527756	786390	67,1	677,7	н.д.
Xeon Phi SE	3775008	5522966	68,4	733,7	н.д.
AMD Radeon HD 7970 (**FirePro W9000**)	106800	226464	47,2	471,1	1884,7
AMD FirePro S10000	421200	1098000	38,4	568,3	2269,44
ATI GPU (**FireStream 9370**)	299300	508499	58,8	310,5	1552,3

** Данные Rmax и Rpeak взяты с http://www.top500.org/statistics/list/*

*** Эффективность Rmax/Rpeak для данного теста взяты с сайта NVIDIA*

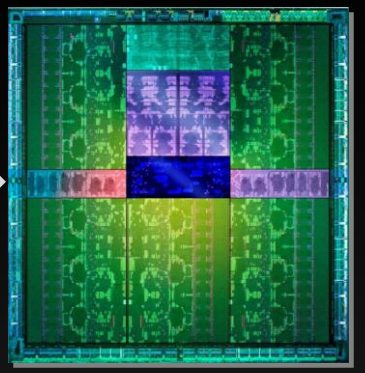

NVIDIA Kepler
Эволюция вычислений на GPU.

Гибридное настоящее и будущее HPC



PCIe

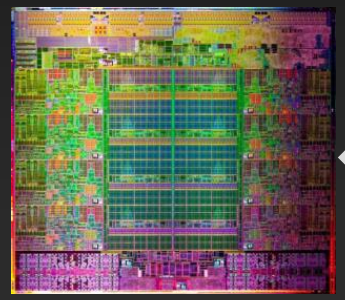
Sandy Bridge
32nm
690 pJ/flop



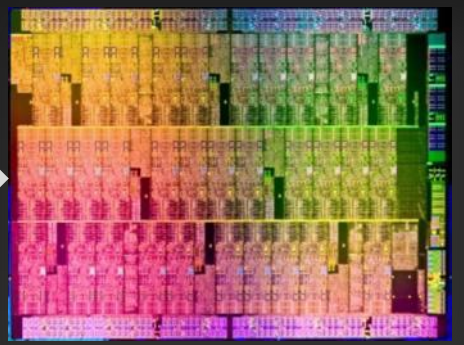
Kepler
28nm
132 pJ/flop

x86 CPU
Быстрая однопоточность
(последовательный код)

GPU
Экстремальная
энергоэффективность
(параллельный код)



PCIe



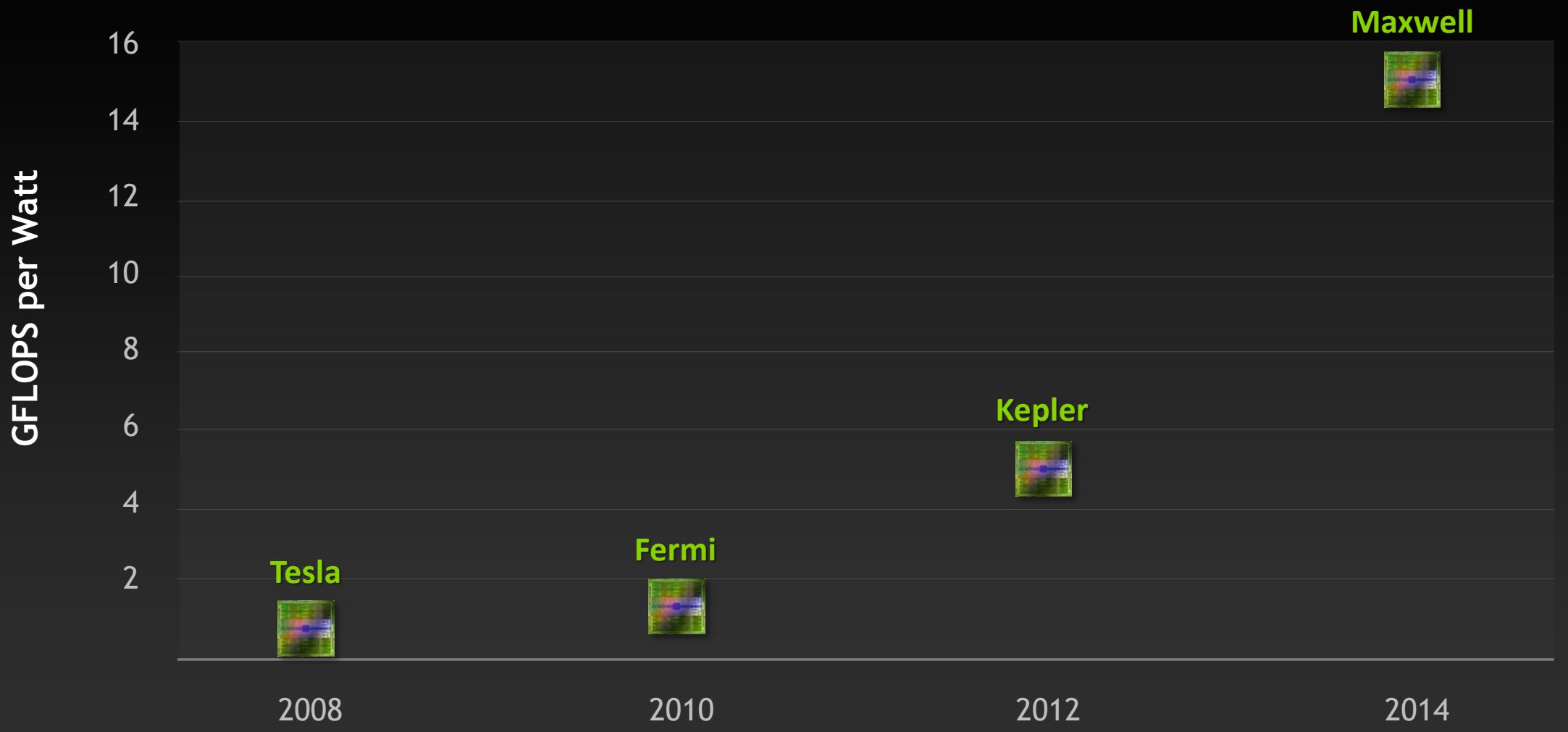
Xeon

Intel MIC

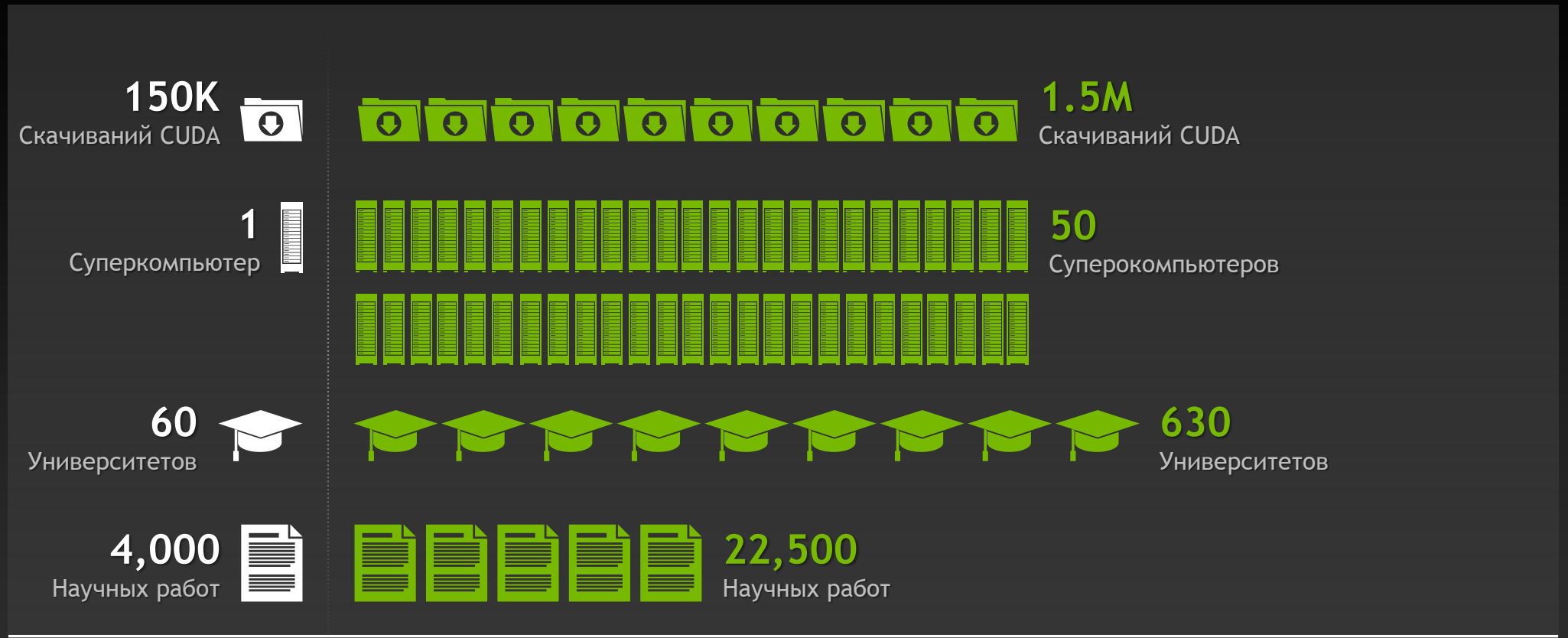
(AMD Fusion)

- Большая часть вычислений на ядрах с **экстремальной энергоэффективностью**
- Несколько ядер для **быстрого выполнения последовательного кода**

Эволюция GPU

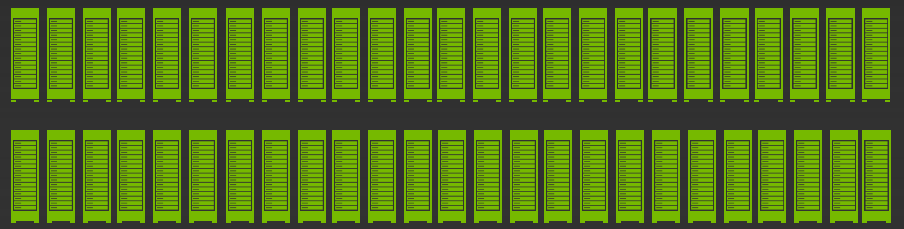


Экспоненциальный рост вычислений на GPU



Вычисления на GPU в России

20M+
GPU с CUDA



15
Суперкомпьютеров

20+
Университетов

350+
Научных работ

2012

Более 200 приложений с поддержкой GPU

http://www.nvidia.com/teslaapps/

POPULAR GPU-ACCELERATED APPLICATIONS

Application	Description	Supported Features	Expected	Multi-GPU	Release Status
Molecular Dynamics					
Abalone	Models molecular dynamics of biopol simulations of proteins, DNA and ligа				
ACEMD	Simulation of mechanics force fields, & explicit solvent on CUDA				
AMBER	Suite of programs to simulate molecu dynamics on biomolecules				
DL-POLY	Simulate macromolecules, polymers, systems, etc on a distributed memory parallel computer				
GROMACS	Simulation of biochemical molecules complicated bond interactions				
HOOMD-Blue	Particle dynamics package written gr up for GPUs				
LAMMPS	Classical molecular dynamics packag				
NAMD	Designed for high-performance simu of large molecular systems				
Quantum Chemistry					
GAMESS-US	Computational chemistry suite used t simulate atomic and molecular electr structure				
NWChem	Computational chemistry package de for HPC clusters				
Q-CHEM	Computational chemistry package de for HPC clusters				
TeraChem	Quantum chemistry software designe to run on NVIDIA GPU				
Materials Science					
LSMS	Materials code for investigating the ef of temperature on magnetism				
QMCPACK	Solves the many-body Schrodinger eq for electronic structures using a quan Monte Carlo method				
Quantum-Espresso/PWscf	An integrated suite of computer code electronic structure calculations and modeling at the nanoscale				
VASP	First principles materials code that c electronic structures and quantum-mechanical molecular dynamics				
Visualization & Docking Software					
Amira 5	A multifaceted software platform for visualizing, manipulating, and unders life sciences and bio-medical data				
Core Hopping	Rapid screening of novel cores to imp drug properties				
FastROCS	3D molecular shape comparison				
VMD	Visualizing and analyzing large bio-m systems in 3-D graphics				

POPULAR GPU-ACCELERATED APPLICATIONS, Continued

Application	Description	Supported Features	Expected	Multi-GPU	Release Status
Weather & Climate Forecasting					
ASUCA	Weather forecasting model fully optimized for GPUs				
CAM / SE	Community Atmospheric Model is a g atmosphere model for weather and c research				
GEOS-5	Weather modeling and forecasting ap used by NASA				
HIRLAM	Weather forecasting model fully optin for GPUs				
HOMME	Weather modeling tool for atmospher scientists				
HYCOM	Weather forecasting model using icos horizontal grid				
MITgcm	Numerical model designed for study atmosphere, ocean, and climate				
NIM	Weather forecasting model using icos horizontal grid				
WRF	Weather and Ocean modeling applica				
Editing and Effects					
Adobe Premiere Pro	Video editing				
Avid Media Composer	Video editing				
GenArts Sapphire	Effects plug-in for video editing				
Sony Vegas Pro	Video editing				
Animation					
Autodesk 3Ds Max	3D modeling, animation, and rendering				
Autodesk Maya	3D modeling, animation, and rendering				
Defense & Intelligence					
DigitalGlobe Advanced Ortho Series	Geospatial Visualization				
Eternix Blaze Terra	Geospatial Visualization				
Exelis (ITT) ENVI	Geospatial Visualization				
GeoEye Analytics Signature Analyst	Geospatial Visualization				
GeoWeb3d Desktop	Geospatial Visualization				
Incogna GIS	Geospatial Visualization				
Intergraph Motion Video Analyst	Video filters and mosaic'ing — Geo-fu FMV analytics with intelligence data				
Intuvision Panoptes 3.0	Video Analytics				

POPULAR GPU-ACCELERATED APPLICATIONS, Continued

Application	Description	Supported Features	Expected	Multi-GPU	Release
Electronic Design Automation and CEM					
Agilent Technologies ADS	Simulation tool for design of RF, micr and high speed digital circuits				
Agilent Technologies EMPro	Modeling and simulation environmen analyzing 3D EM effects of high speed RF/Microwave components				
ANSYS Nexxim	Circuit simulation engine for RF/anal mixed-signal IC design; IBIS-AMI ana speedup with GPU computing				
CST Microwave Studio (MWS)	High frequency electromagnetic field simulation				
Gauda OPC, OPV	Collection of several software tools fo computational lithography running or Gauda hardware platform				
Remcom XFdtd	3D EM modeling and simulation				
Rocketick RocketSim	Verilog simulation				
SPEAG SEMCAD-X	3D EM modeling and simulation				
CAD					
CATIA V6 – Live Rendering	Photorealistic rendering				
Bunkspeed Pro Suite	Easy to use photorealistic rendering s				
RTT DeltaGen 10.x	Photorealistic rendering used for des				
RTT DeltaPix	Photorealistic rendering with integrat TeamCenter and RTT formats				
Numerical Analysis					
Jacket AccelerEyes	GPU acceleration for MATLAB				
Mathematica Wolfram	Symbolic math analysis				
MATLAB Mathworks	Technical computing language and integrated development environment (MATLAB PCT, MDCS)				

POPULAR GPU-ACCELERATED APPLICATIONS, Continued

Application	Description	Supported Features	Expected	Multi-GPU	Release
Oil&Gas					
Acceleware RTM	Seismic Processing				
CGG/Veritas RTM	Seismic Processing				
ffA SVI Pro	Seismic Interpretation				
Headwave Suite	Seismic Imaging				
Geoteric	Seismic Processing/Interpretation				
Paradigm EarthStudy360	Reservoir Modeling				
Paradigm Echos RTM	Seismic Processing				
Paradigm SKUA	Reservoir Modeling				
Paradigm VoxelGeo	Seismic Interpretation				
Schlumberger WesterGeco Omega2 RTM	Seismic Processing				
Seismic City Prestack Interpretation	Seismic Processing				
SpectraSeis	Seismic Processing / Imaging				
Stoneridge Reservoir Simulation	Reservoir Simulation				
Tsunami RTM	Seismic Processing				
Computational Finance					
Hanweck Associates	Real-time options analytical engine [				
MATLAB Mathworks	Data parallel mathematics (MATLAB PCT, MDCS)				
Murex	Risk analytics (MACS)				
Numerical Algorithms Group	Random Number Generators				
SciComp, Inc	Derivative pricing (SciFinance)				
Wolfram Mathematica	Mathematical Development Environ				

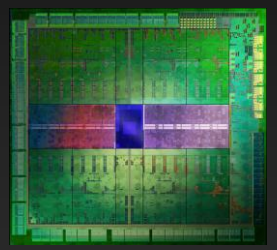
*GPU performance compared against multi-core x86 CPU socket. kernel to kernel performance comparison. Performance results

POPULAR GPU-ACCELERATED APPLICATIONS, Continued

Application	Description	Supported Features	Expected Speed Up*	Multi-GPU Support	Release Status
Physics					
Chroma	General purpose LQCD application	Wilson-clover fermions, Krylov solvers, Domain-decomposition	5-6x	Yes	Available now
MILC	General purpose LQCD application	Staggered fermions, Krylov solvers, Gauge-link fattening	5-6x	Yes	Available now
Computational Fluid Dynamics					
Altair AcuSolve	General purpose CFD flow solver	Linear equation solver	2x	Yes	Available now
Autodesk Moldflow	Optimize design of plastic parts and injection molds	Linear equation solver	2x	Single Only	Available now
FEFLO (GMU-Lohner)	Navier-Stokes flow solver based on unstructured grids for modeling both compressible and incompressible flows	Explicit solver	10x	Yes	In Development
FluiDyna LBultra	Computing physical flows in and around solid bodies	LBM, particle CFD	20x	Yes	Available now
FluiDyna Culises-OpenFOAM	Computing physical flows with Culises — a software library with special algorithms for solving systems of equations	Linear equation solvers	3x Solver	Single Only	Available now
Promotech Particleworks	Fluid simulation for free surface flow like Tsunami, material processing and liquids	MPS, Particle CFD	4x-9x	Yes	Available now
S3D (Sandia NL S3D)	Massively parallel direct numerical solver (DNS) for the full compressible Navier-Stokes	Chemistry kernel	8x SP, 5x DP kernel	Yes	In Development
Turbostream	Ultrafast CFD solver for turbomachines	Explicit solver	19x	Yes	Available now
Vratis SpeedIT-OpenFOAM Solver	Set of accelerated solvers for sparse linear systems of equations	Linear equation solvers	3x Solver	Yes	Available now
Computational Structural Mechanics					
Abaqus/Standard	Simulation and analysis tool for structural mechanics	Linear equation solver	1.5-2.5x	Single Only	Available now
ANSYS Mechanical	Simulation and analysis tool for structural mechanics	Linear equation solver	2x	Single Only	Available now
Impetus Afea	Predicts large deformations of structures and components exposed to extreme loading conditions	Linear equation solver, SPH	10x SPH, 2x Total	Yes	Available now
LS-DYNA Implicit	Multiphysics simulation package used	Linear equation solver	3x	Yes	In Development
MSC Nastran	Simulation and analysis tool for structural mechanics	Linear equation solver	1.4-2x	Yes	Available now
Marc	Simulation and analysis tool for structural mechanics	Linear equation solver	1.5x	Yes	In Development
RADIOSS Implicit	Used to maximize durability, NVH, crash, safety, manufacturability and fluid-structure interaction performance	Linear equation solver	2x	Single Only	In Development

Продукты Tesla на базе архитектуры Kepler

Tesla K10

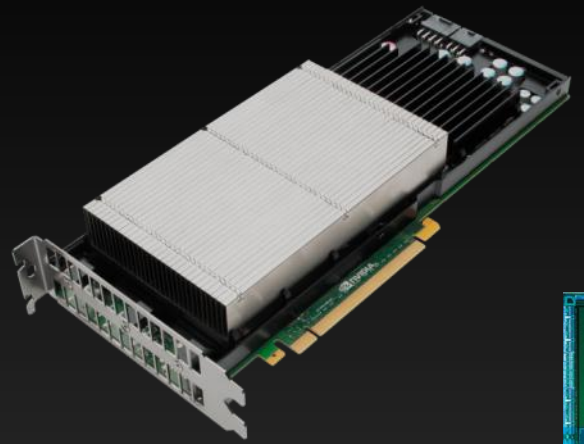


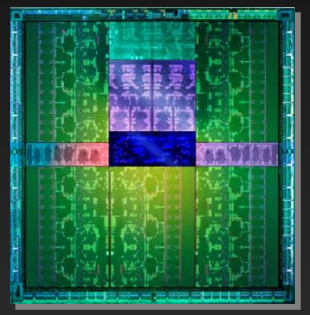
Два GK104 GPU

3x одинарная точность

Обработка видео, сигналов, сейсмика, молекулярная динамика

Tesla K20





GK110 GPU

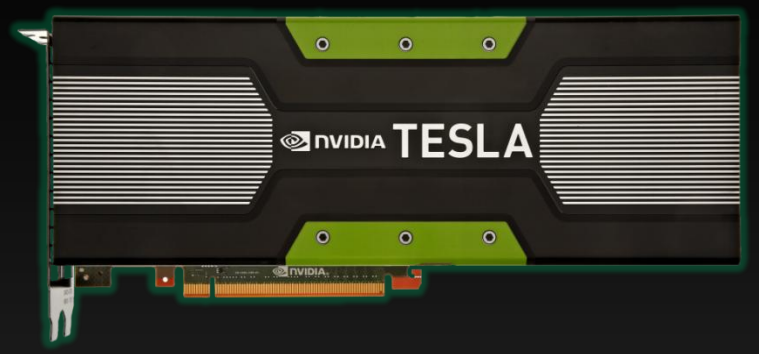
3x двойная точность

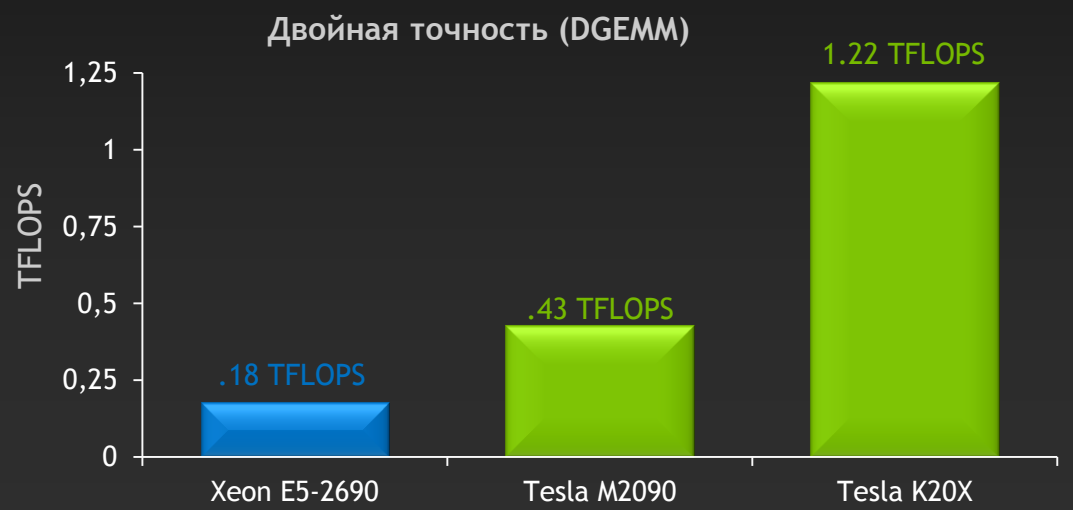
Гидро-, газо- динамика, прочностной анализ, финансы, физика и пр.

Уже в продаже

Семейство Tesla K20 : в 3 раза быстрее Fermi

Tesla K20X



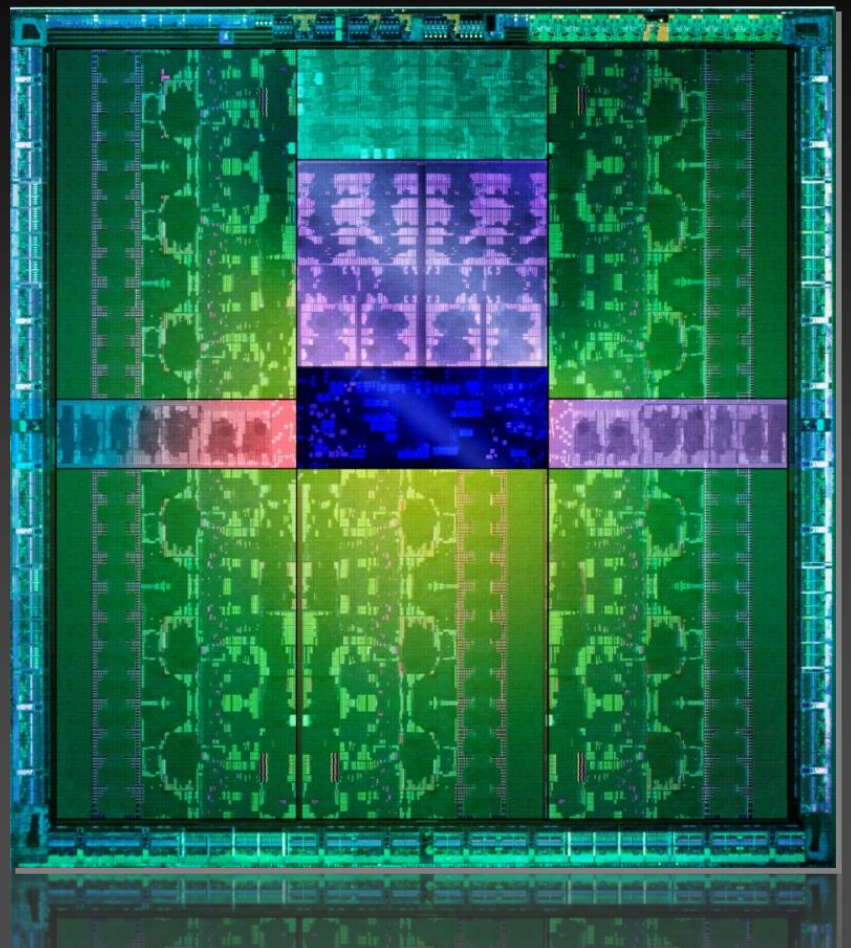


	Tesla K20X	Tesla K20
Кол-во CUDA ядер	2688	2496
Пиковая произв-ть DP DGEMM	1.32 TF 1.22 TF	1.17 TF 1.10 TF
Пиковая произв-ть SP SGEMM	3.95 TF 2.90 TF	3.52 TF 2.61 TF
Пропускная способность памяти	250 GB/s	208 GB/s
Объем памяти	6 GB	5 GB
Потребление	235W	225W

Kepler

САМАЯ БЫСТРАЯ И ЭФФЕКТИВНАЯ HPC АРХИТЕКТУРА

GK110 GPU



SMX *(энергоэффективность)*

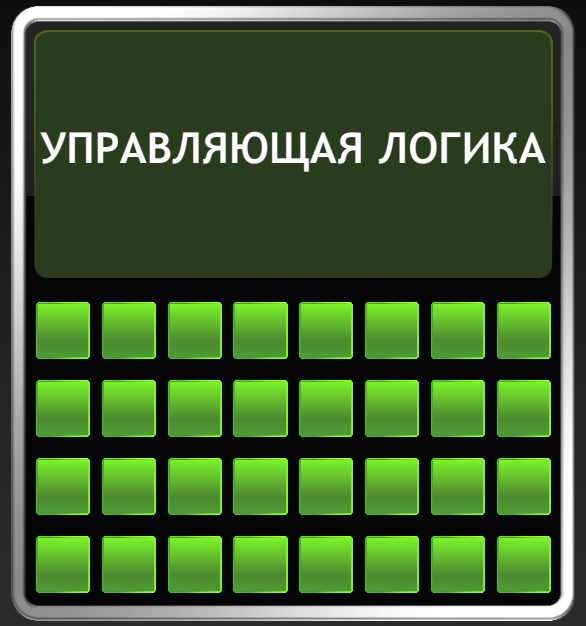
Hyper-Q *(удобство программирования и применимость)*

Dynamic Parallelism

Kepler: производительность и эффективность

SM

M2090



32 ядра

3x

Perf / Watt

SMX

K20

192 ядра

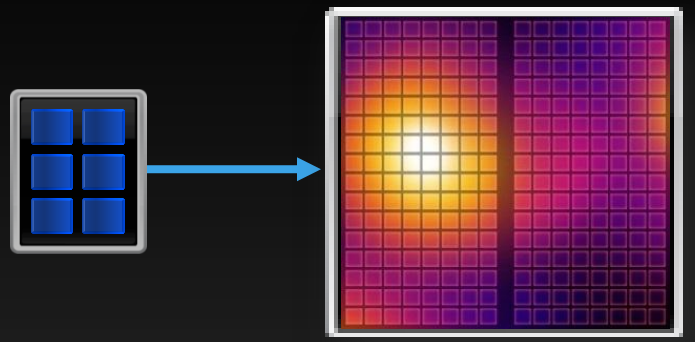

1 Петафлопс
Всего в 10 стойках
400 кВт

Hyper-Q

Простота использования с MPI приложениями

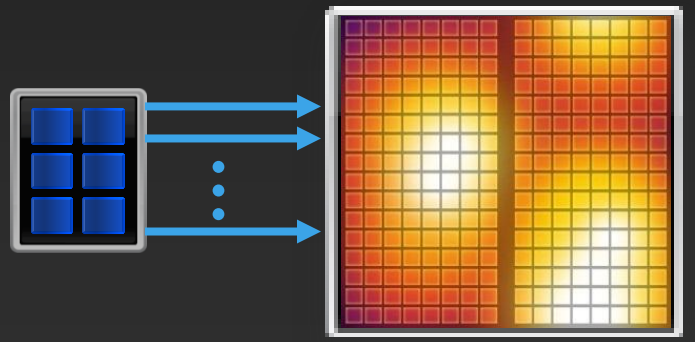
FERMI

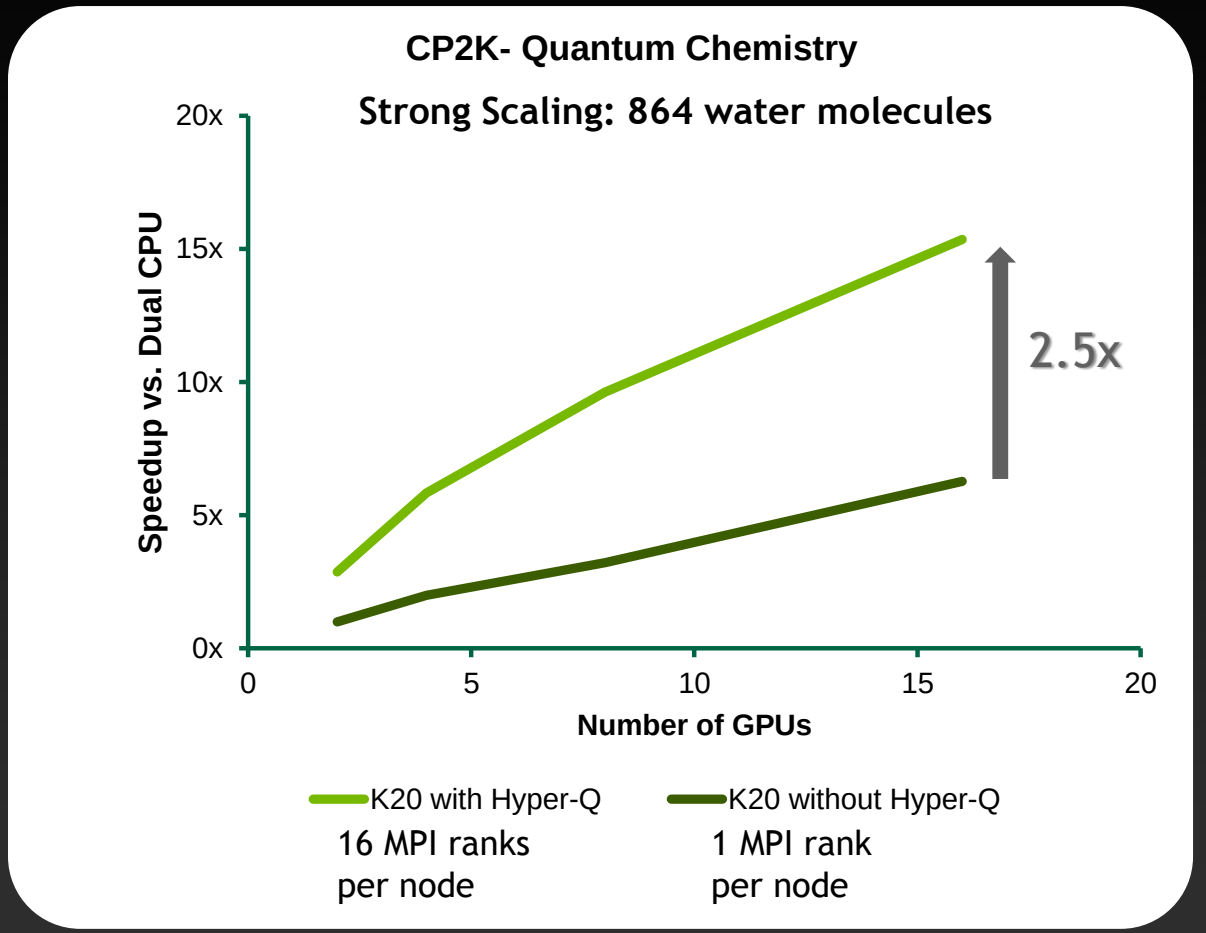
1 очередь задач



KEPLER

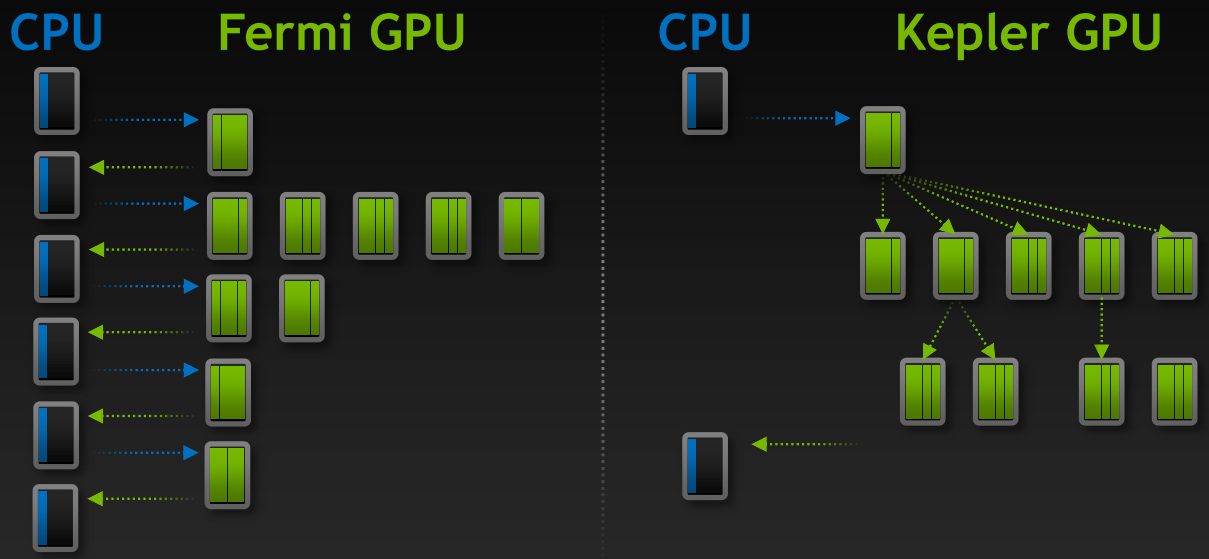
32 очереди задач

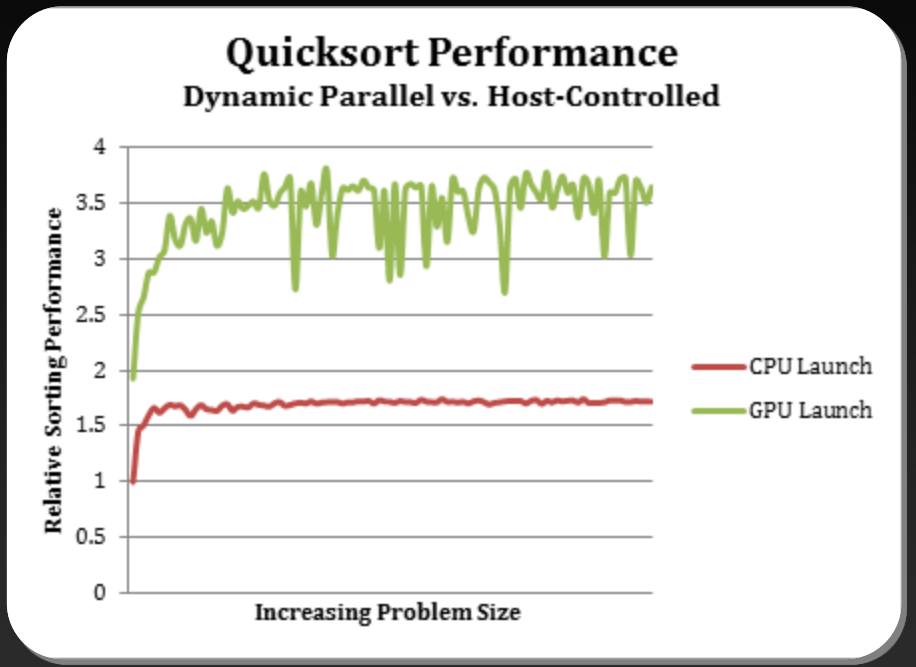




Dynamic Parallelism

Проще и понятнее код, выше производительность





Код в 2 раза короче

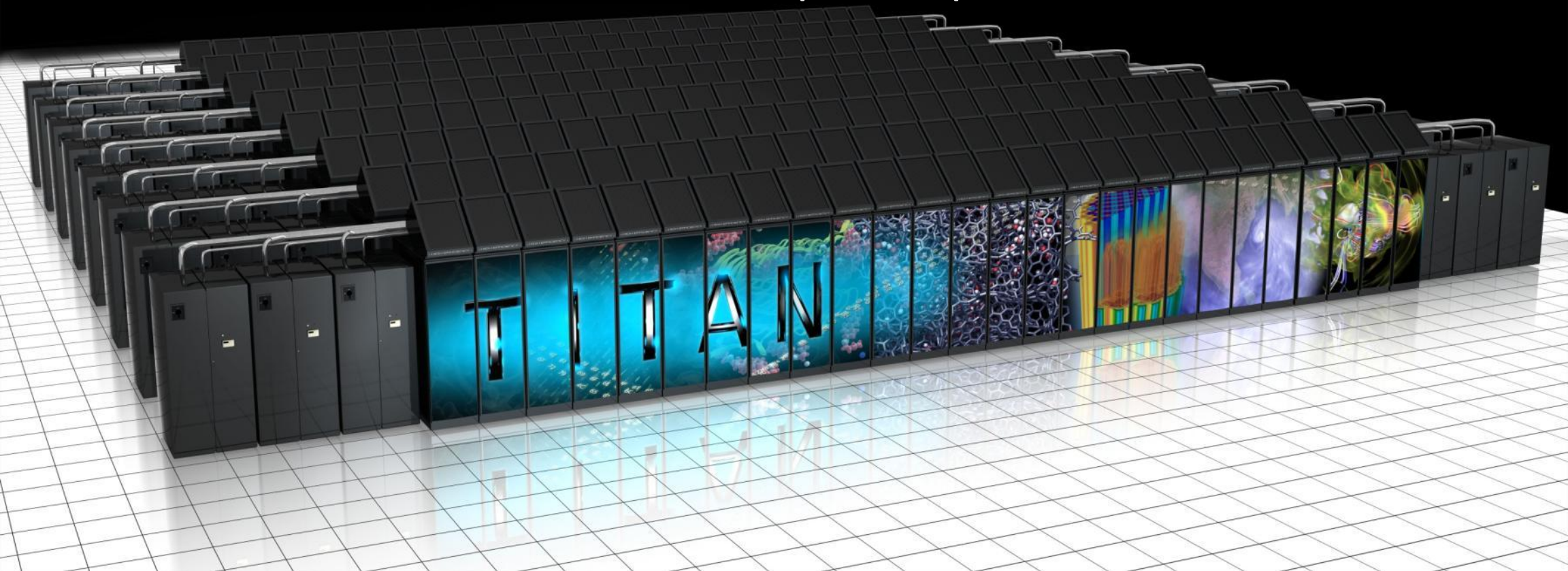
Производительность в 2 раза выше

Titan: самый быстрый суперкомпьютер в мире

18,688 Tesla K20X GPU

27 Petaflops пик: 90% производительности обеспечено GPU

17.59 Petaflops в Linpack



На 60% больше CUDA приложений, включая все ключевые

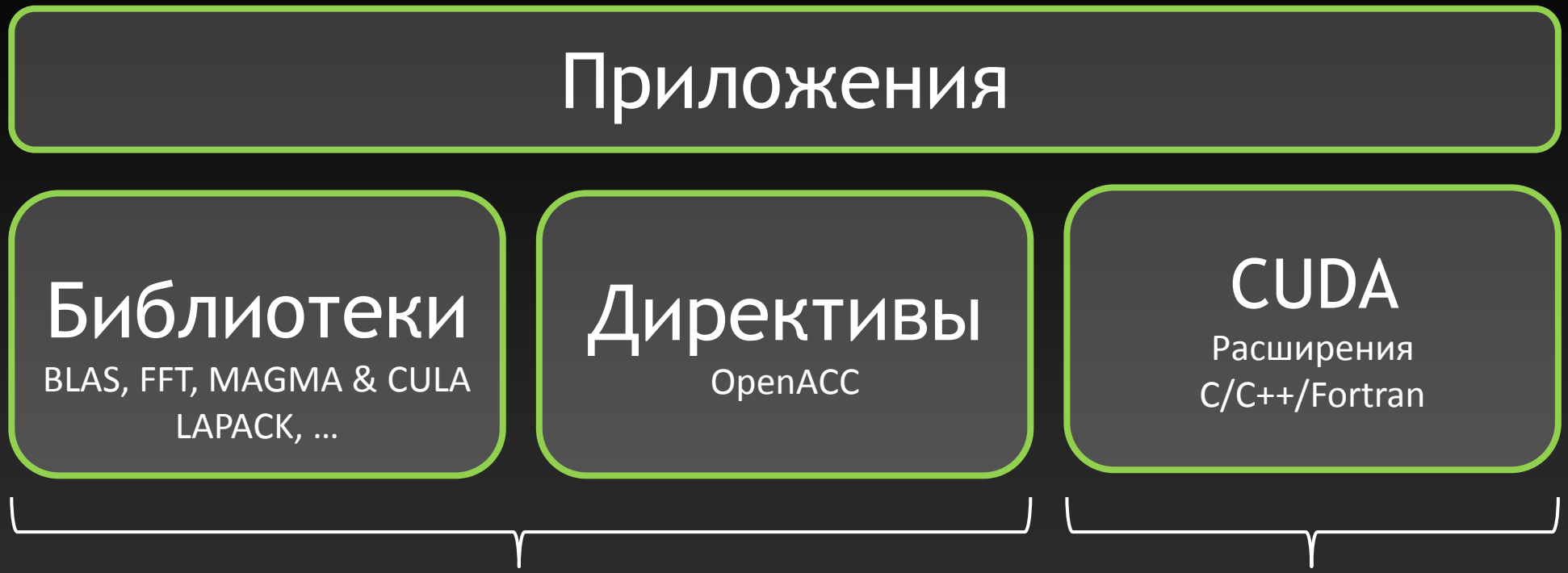
Топовые приложения для СК

Вычислительная химия	AMBER CHARMM GROMACS	LAMMPS NAMD DL_POLY
Материаловедение	QMCPACK Quantum Espresso GAMESS	Gaussian NWChem VASP
Климат и погода	COSMO GEOS-5	CAM-SE NIM WRF
Физика	Chroma Denovo GTC	GTS ENZO MILC
CAE	ANSYS Mechanical MSC Nastran SIMULIA Abaqus	ANSYS Fluent OpenFOAM LS-DYNA

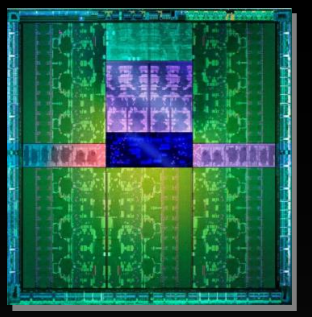
Каталог приложений: http://www.nvidia.com/teslaapps/

Оптимизированы, в разработке

Программирование на GPU



Выводы

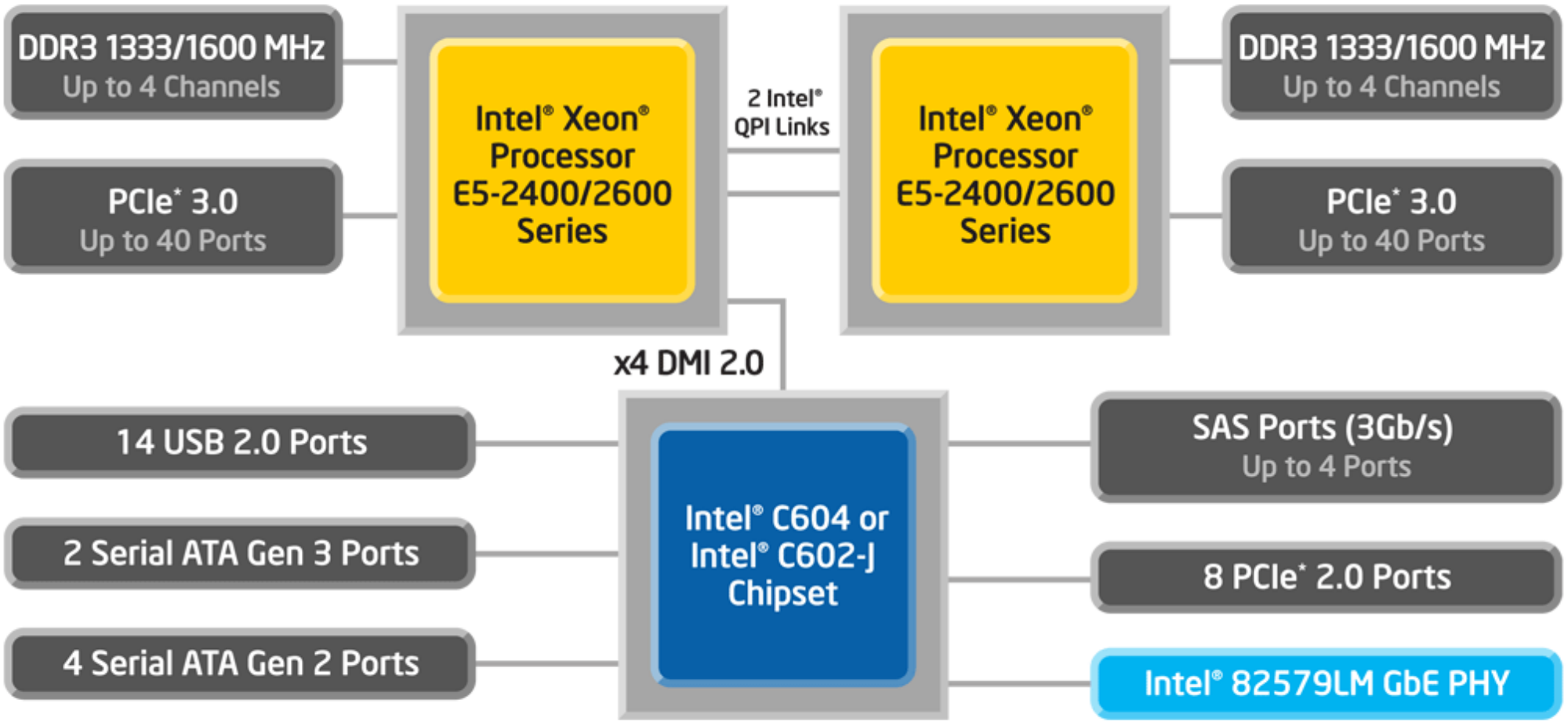


- Вычисления на GPU ускорителях *значительно эволюционировали за очень короткое время*
  - Стали удобнее для программирования и полностью универсальны
- Вычисления на GPU базируются на устойчивой бизнес модели
  - Находятся в технологическом авангарде с поддержкой со стороны массового рынка
- Вычисления на ускорителях теперь признаны всеми игроками рынка как правильное направление развития
  - HPC процессоры будущего будут строиться на гибридной архитектуре
- Kepler – самая быстрая и эффективная HPC архитектура

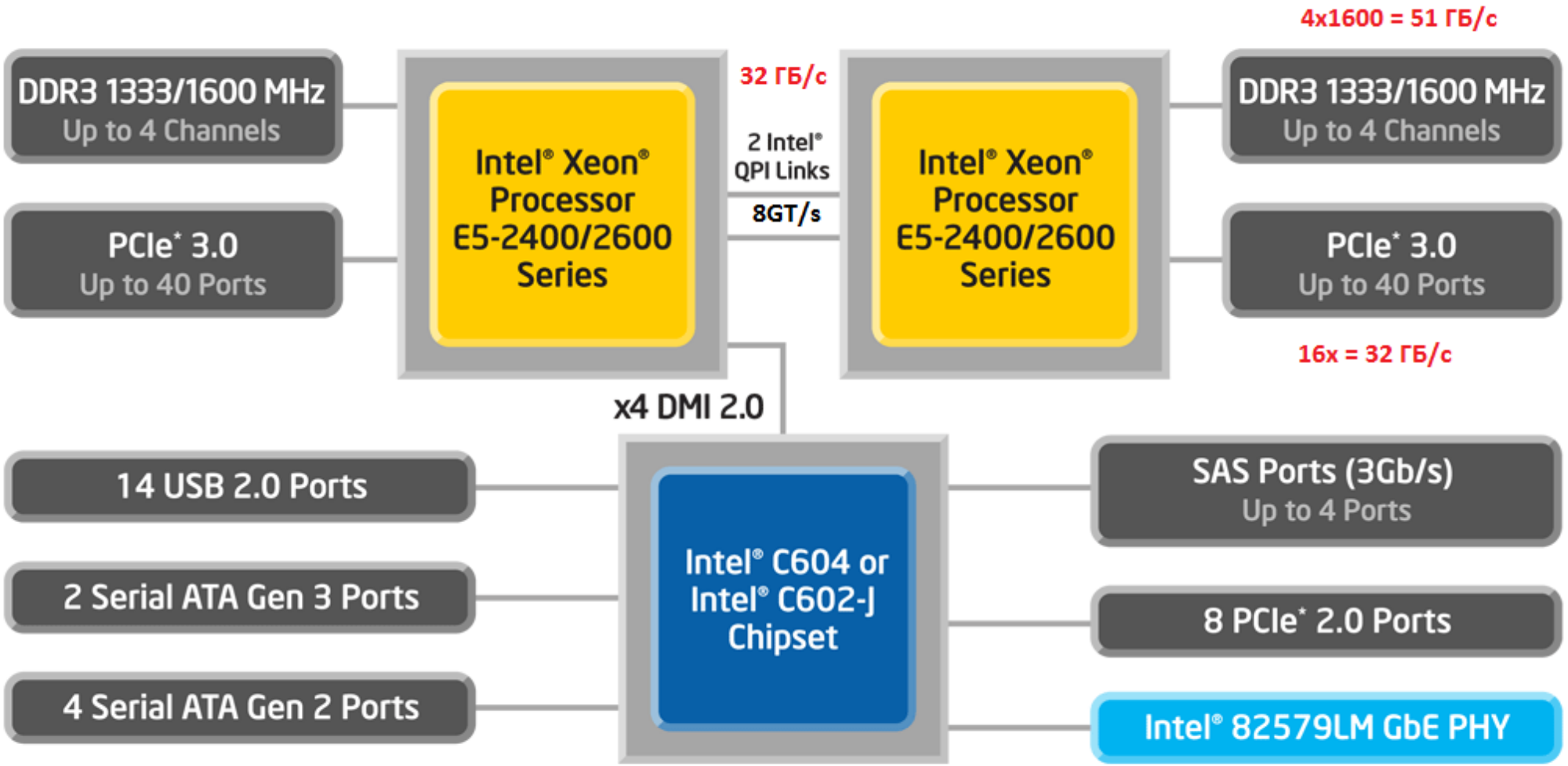
Выбор CPU

- Чипсет – важно ли?
- Баланс CPU и GPU
- CPU – 1, 2, 4 или 8 на ноду?
- AMD или Intel?

Чипсет – важно ли?



Чипсет – важно ли?

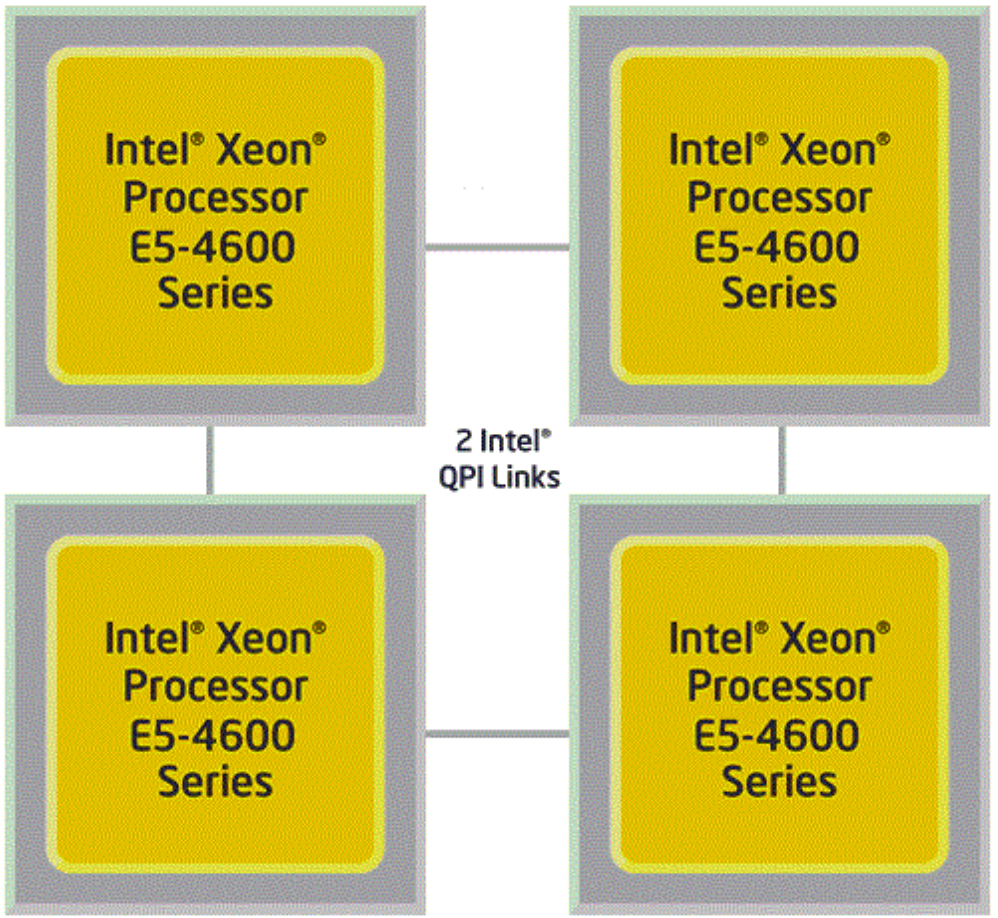


Чипсет – важно ли?

Выводы:

- Самым слабым местом является шина процессора
- Минимальным рекомендуемым процессором является Intel Xeon Processor E5-2650 (20M Cache, 2.00 GHz, 8.00 GT/s QPI)
- Оптимальным с учетом цена/производительность – Intel Xeon Processor E5-2670 (20M Cache, 2.60 GHz, 8.00 GT/s QPI)

CPU – 1, 2, 4 или 8 на ноду?



AMD или Intel?

Model	SPECint®_ rate20063	SPECfp®_ rate2006	SPECint®_ rate_base2006	GFLOPs (Theoretical)	Total Processor 1kU Price - 2P*
Intel Xeon™ processor Model E5-2690	705	510	671	371	$4,114
Intel Xeon™ processor Model E5-2670	647	486	622	333	$3,104
Intel Xeon™ processor Model E5-2660	596	459	572	282	$2,658
AMD Opteron™ processor Model 6284 SE	573	417	499	346	$2,530
AMD Opteron™ processor Model 6282 SE	543	403	474	333	$2,038
Intel Xeon™ processor Model E5-2650	543	433	521	256	$2,214

Взято с сайта: http://sites.amd.com/us/Documents/49747D_HPC_Processor_Comparison_v3_July2012.pdf

Производительность и цена серий Intel E5-2600 и AMD Opteron 6200 сопоставима, но топовые E5 быстрее топовых Opteron

AMD или Intel?

Где может быть лучше AMD:

- В 4-х процессорных нодах – 4 QPI 6,4 GT/s против 2 QPI 8 GT/s у Intel
- В 1-процессорных блейдах, где нужно много физических ядер
- При большом бюджете – топовые решения AMD как правило дешевле топовых решений Intel (не наш случай)

Немного теории

Скорость расчета параллельной задачи на кластере определяется формулой:

$$T = \frac{T_c + T_i}{n}$$

При количестве нод в кластере больше 8, время на обмен информацией между нодами в разы больше, чем время расчета одной итерации процессором (CPU или GPU)

Вывод: при выборе любого компонента в первую очередь всегда смотрим на скорость его шины

Интерконнект – Ethernet или InfiniBand?

	1Gb Ethernet	10 Gb Ethernet	QDR	FDR	EDR
Пропускная способность интерфейса, raw / data, Гбит/с	1 / 0,8	10 / 8	40 / 32	56,25 / 56	103,12 / 100
Латентность, мкс	100	10	3	1,2	<1
RDMA	нет	нет	да	да	да
Цена двухпортового адаптера, $	300	600	1000	1300	н.д.
Обязателен коммутатор, $+	да	да	нет	нет	нет

Распространенные заблуждения:

- InfiniBand – дорого
- InfiniBand не намного быстрее
- InfiniBand сложен в настройке

Интерконнект – Ethernet или InfiniBand?

	1Gb Ethernet	10 Gb Ethernet	QDR	FDR	EDR
Пропускная способность интерфейса, raw / data, Гбит/с	1 / 0,8	10 / 8	40 / 32	56,25 / 56	103,12 / 100
Латентность, мкс	100	10	3	1,2	<1
RDMA	нет	нет	да	да	да
Цена двухпортового адаптера, $	300	600	1000	1300	н.д.
Обязателен коммутатор, $+	да	да	нет	нет	нет

Распространенные заблуждения:

- InfiniBand – ~~дорого~~ дешевле чем Ethernet
- InfiniBand ~~не намного быстрее~~ на порядок быстрее
- InfiniBand ~~сложен в настройке~~ проще в настройке

Отличительные особенности FDR

Улучшения FDR InfiniBand в сравнении с QDR:

- Параметры Link speed увеличились до 14 Гбит/с на линию или 56 Гбит/с по четырем линиям
- Показатель Link кодировки для FDR InfiniBand был изменен с 8 бит/10 бит на 64 бит/66 бит
- Улучшены механизмы коррекции ошибок сети
- Реальное удвоение производительности – на сегодня это самый производительный интерконнект и в абсолютном значении, и в пересчете на $

Много слабых нод или мало, но мощных?

Увлекаться количеством узлов не стоит

Самое узкое место в вашем кластере - это среда передачи данных между узлами, то есть пропускная способность используемой сети

Итоги

Даже в рамах сильно ограниченного бюджета можно собрать кластер, обладающий хорошей производительностью

Каждая нода кластера состоит из:

- 2 процессоров Intel Xeon Processor E5-2670 (20M Cache, 2.60 GHz, 8.00 GT/s QPI);
- 2 GPU-ускорителей Tesla на архитектуре Kepler.

Между собой ноды объединены InfiniBand FDR от Mellanox

TECHNICAL SUPERCOMPUTERS | LABORATORY

Вывод

Для построения сбалансированного кластера нужно обратиться к нам

ООО «Терминал-Сервис»,

Сергей Дудинов,

Директор

sd@tslab.com.ua