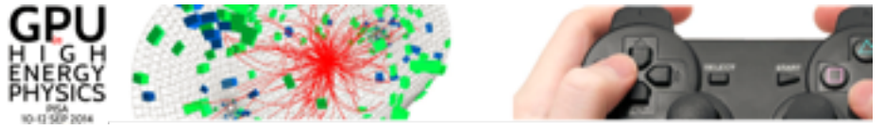
Conferenze su GPGPU in Fisica a Pisa e Roma: Florilegio dall'avamposto CNAF



D. Cesini, L. Morganti, S. A. Tupputi

perché

The hardware producers commitment in the market of computer graphics and games together with the on-going development of proprietary and free software expose the raw computing power of GPUs for general-purpose applications and scientific computing in particular.

- Pisa

The use of graphic processors as accelerators in data- and computation-intensive applications found fertile ground in the HEP community and is currently object of active investigations.

Computational speed-ups in online and offline data selection and analysis to hard real-time applications in low-level triggering, to MonteCarlo simulations for lattice QCD.

- Roma

The aim of the meeting is to present and discuss some of the modern applications in Physics and Astrophysics (and related subfields) of hybrid computational systems based on multicore CPU governing a set of Graphic Processing Units (GPUs) acting as number crunchers.

Menu variegato: cosmologia, sismologia, biologia molecolare, machine learning, HEP (of course), astrofisica

compendio

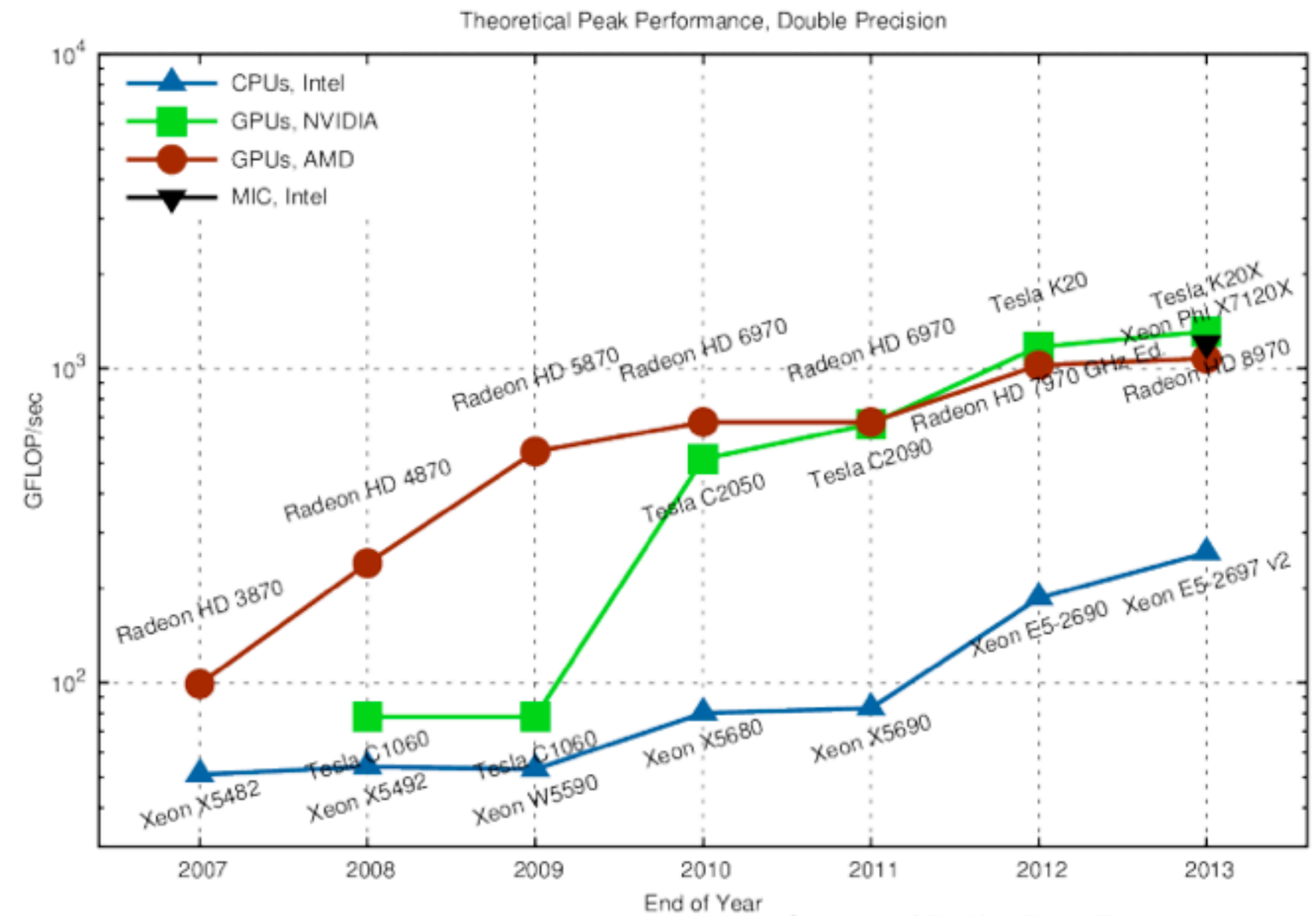
- Trend e progetti

- Aree tematiche
    - HEP - Astrofisica - Fisica Teorica - Altro (terremoti, meteorologia, dinamica molecolare)
    - Progetti multi disciplinari
- Piattaforme e linguaggi
    - NVIDIA vs INTEL, HW & SW...

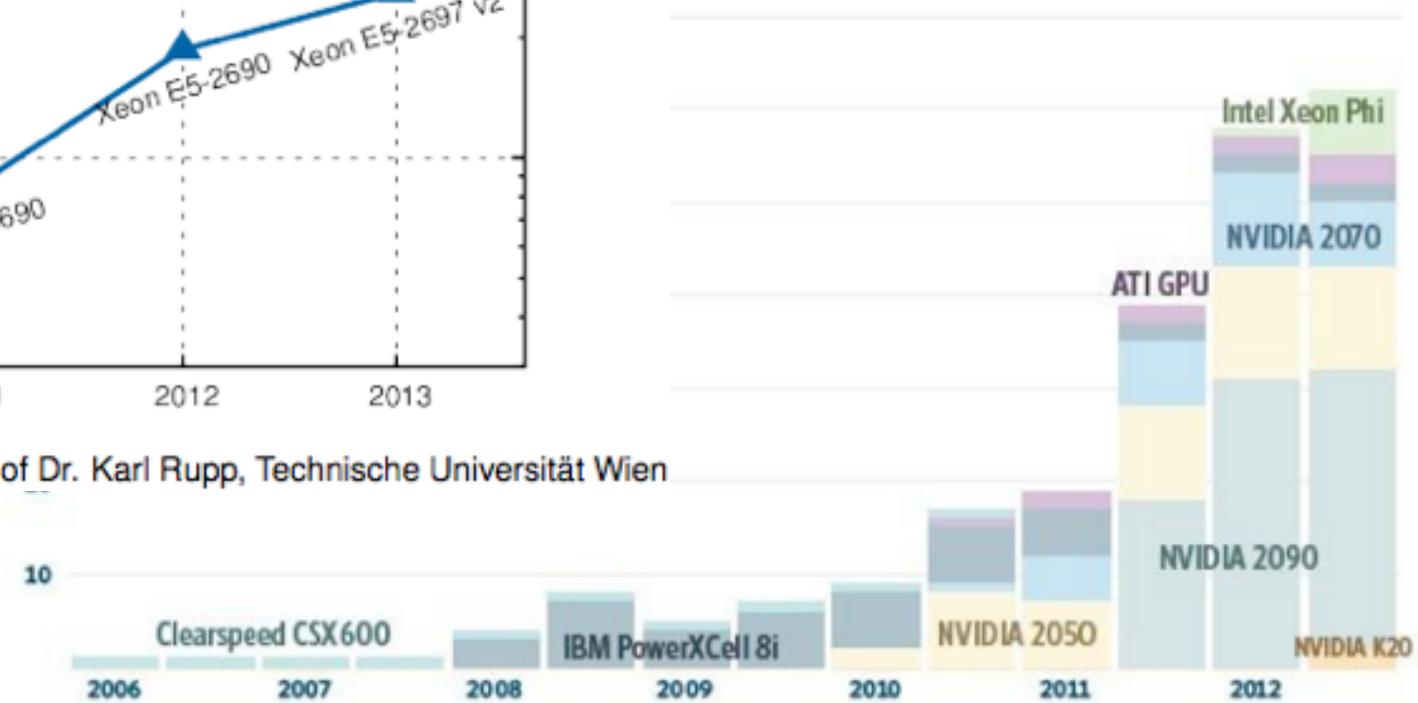
disclaimer (da Daniele)

- Il livello medio dei numeri presentati nel confronto CPU/GPU lascia perpelssi
  - codice seriale vs codice CUDA
  - slide e presentazioni priviliegiavano l'esposizione del problema rispetto alla descrizione dei dispositivi usati
- speedup poco credibili
  - vedi: "Twelve ways to fool the masses with speedups"

Trends



Courtesy of Dr. Karl Rupp, Technische Universität Wien



Accelerator architectures in the Top500 Supercomputers

Languages & libraries

- OpenCL: code and memory movements explicitly programmed – on COU, GOU, MIC, ...
- OpenACC/OpenMP: offloads functions with #pragma and data movements may be manually managed
- CUDA
  - possibile compilare per C++11
  - Thrust: STL-like CUDA template library
  - CUB: CUDA Unbound: building blocks for kernels via C++library
  - CudNN: deep neural network library (pre-packaged kernels)
  - Non supporta piu OpenCL, spinge invece per OpenACC

Alcune idee condivise

- HW: BE-connected intermediate layer for abstracting from HW architectures and purpose of object structures for mapping suited device to suited algorithm

- SW
  - exploit parallelism at different levels
  - Across different architectures
  - No perf degradation when ported

Suite - Tools - Framework

- GWTools: modular building blocks for various applications to be run on GPU, CPU, APU, hybrid systems
  - computer wrapper with multiple backedns (CUDA ,OpenCL, ...)
  - claimed to be highly optimized
- GeantV introduce nel pacchetto tecniche HPC (vettorizzazione, istruzioni multiple, GPGPU)
  - raccordo tra algoritmi e HW tramite librerie e opportuni BE
- Parecchi “pacchetti” di calcolo per LQCD

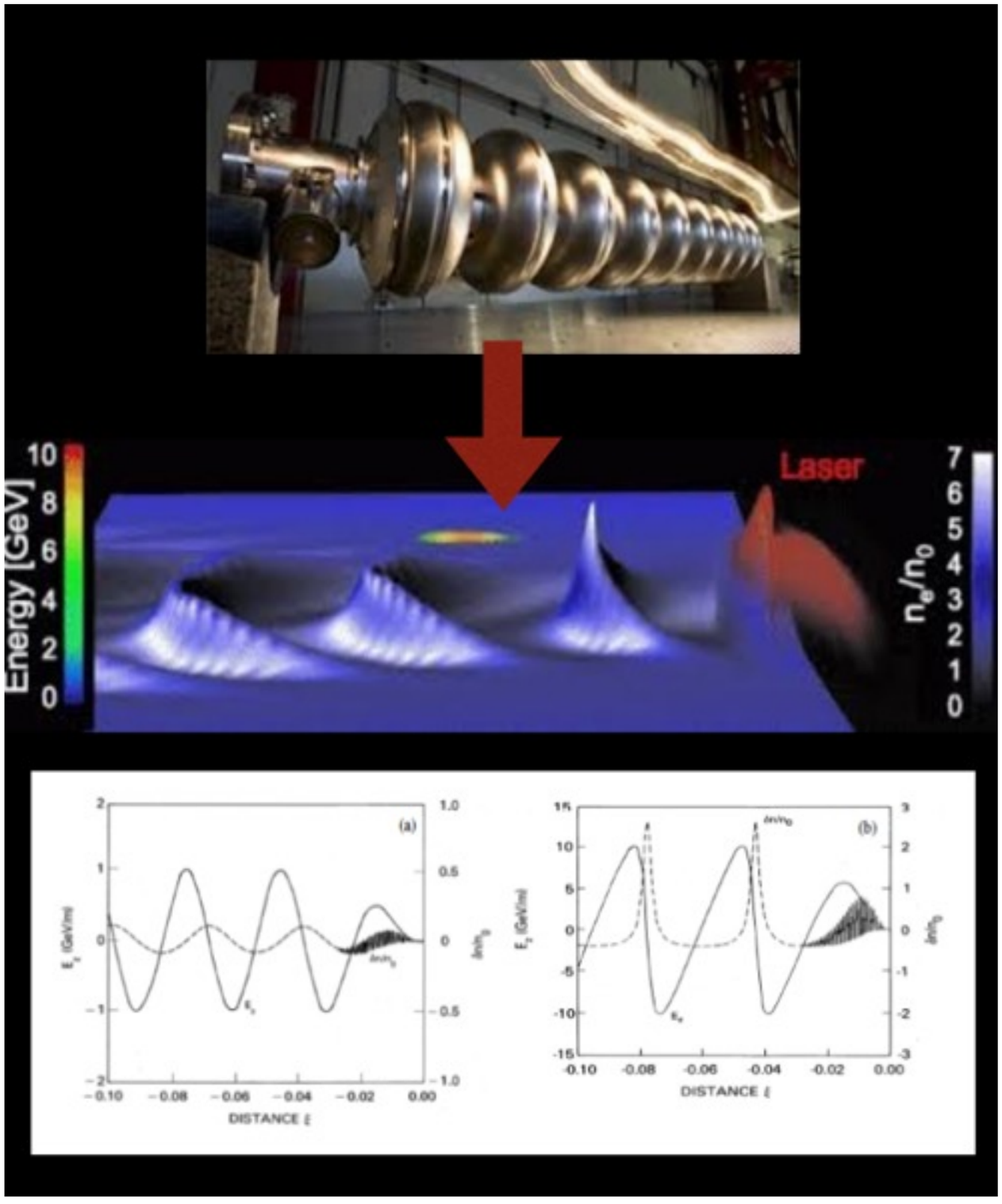
CNAF alle conferenze

- F. Rossi: Robust algorithms for current deposition and dynamic load-balancing in a GPU particle-in-cell code

- APE cluster

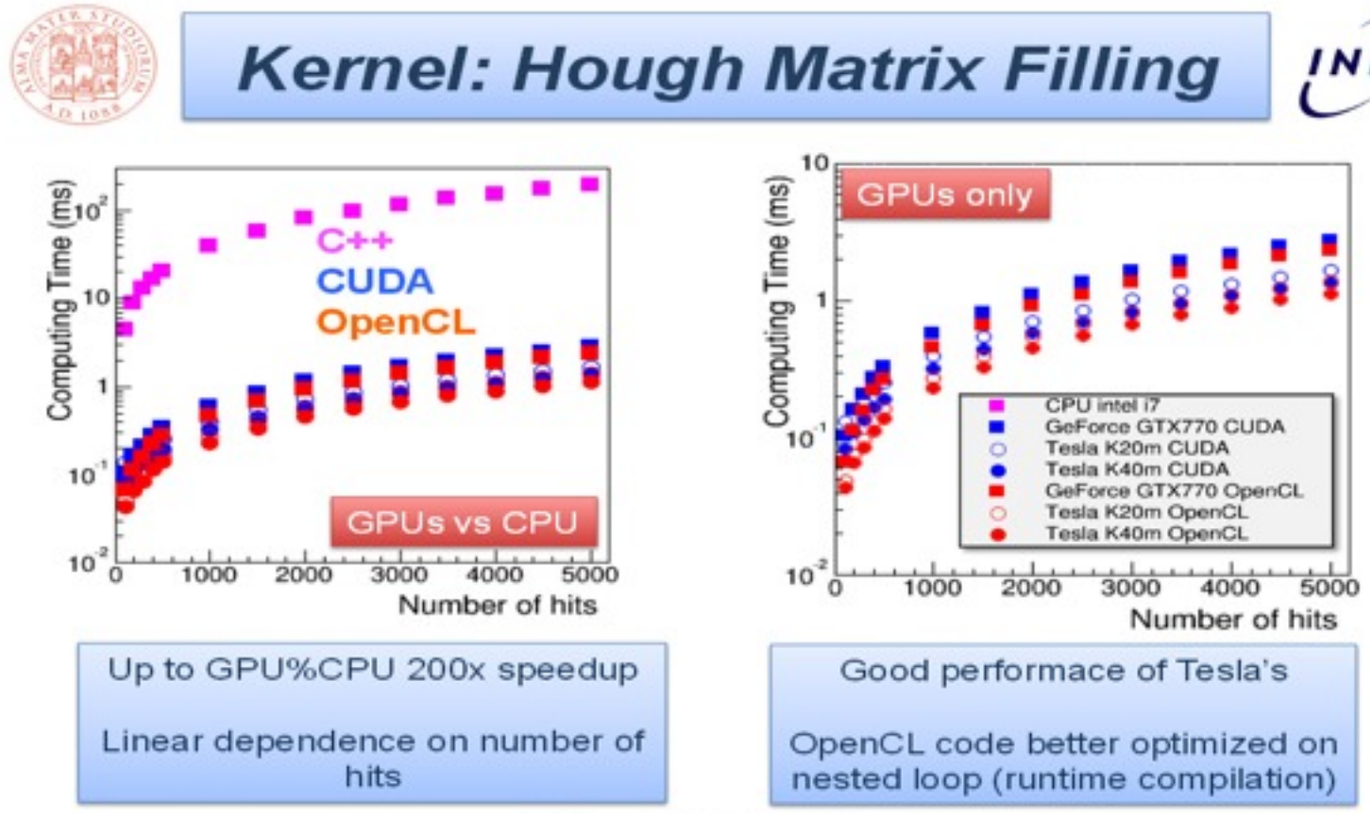
- CINECA

- Cluster del CNAF



CNAF alle conferenze

- M. Belgiovine, L. Rinaldi, S. A. Tupputi et al. - GPGPU for track finding in HEP



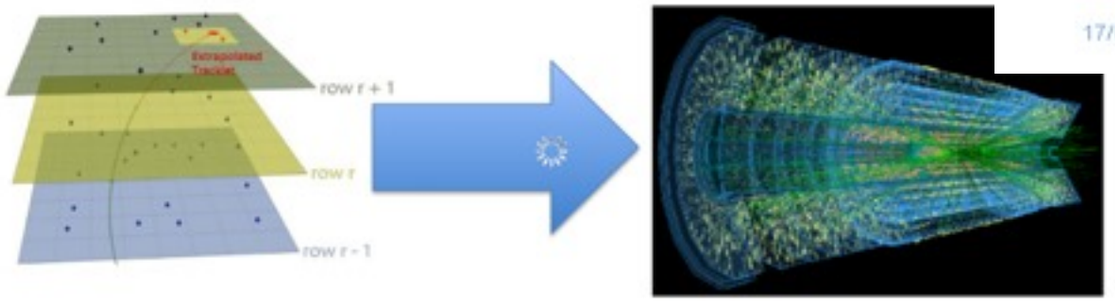
Up to GPU%CPU 200x speedup

Linear dependence on number of hits

Good performace of Tesla's

OpenCL code better optimized on nested loop (runtime compilation)

A massive parallel approach based on GPGPU can be relevar Tracking in High Energy Physics



Fast tracking is suitable for realtime data selection

In this contribution we will show a track finding algorithm based on the Hough Transform

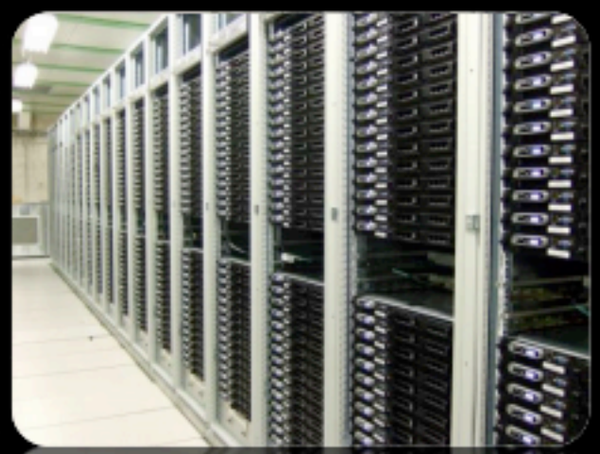
HEP

COMPUTATIONAL CHALLENGES IN HEP

Low-Level Trigger

Connectivity
Latency control
Power limited

High-Level Trigger

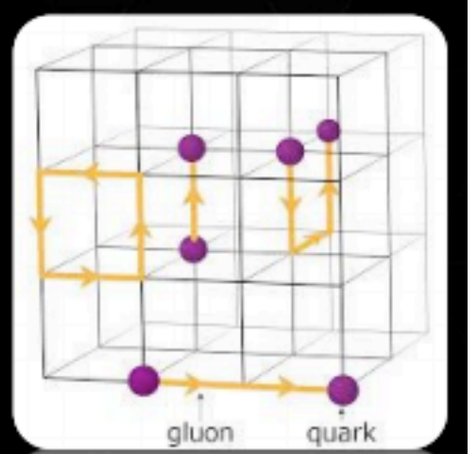


Data volume
Computer vision
Unsupervised learning

Monte Carlo Analysis

Portability
Legacy codes
Limited parallelism
Geometry processing

Lattice QCD



Connectivity
Low latency
Bandwidth,
bandwidth,
bandwidth,..

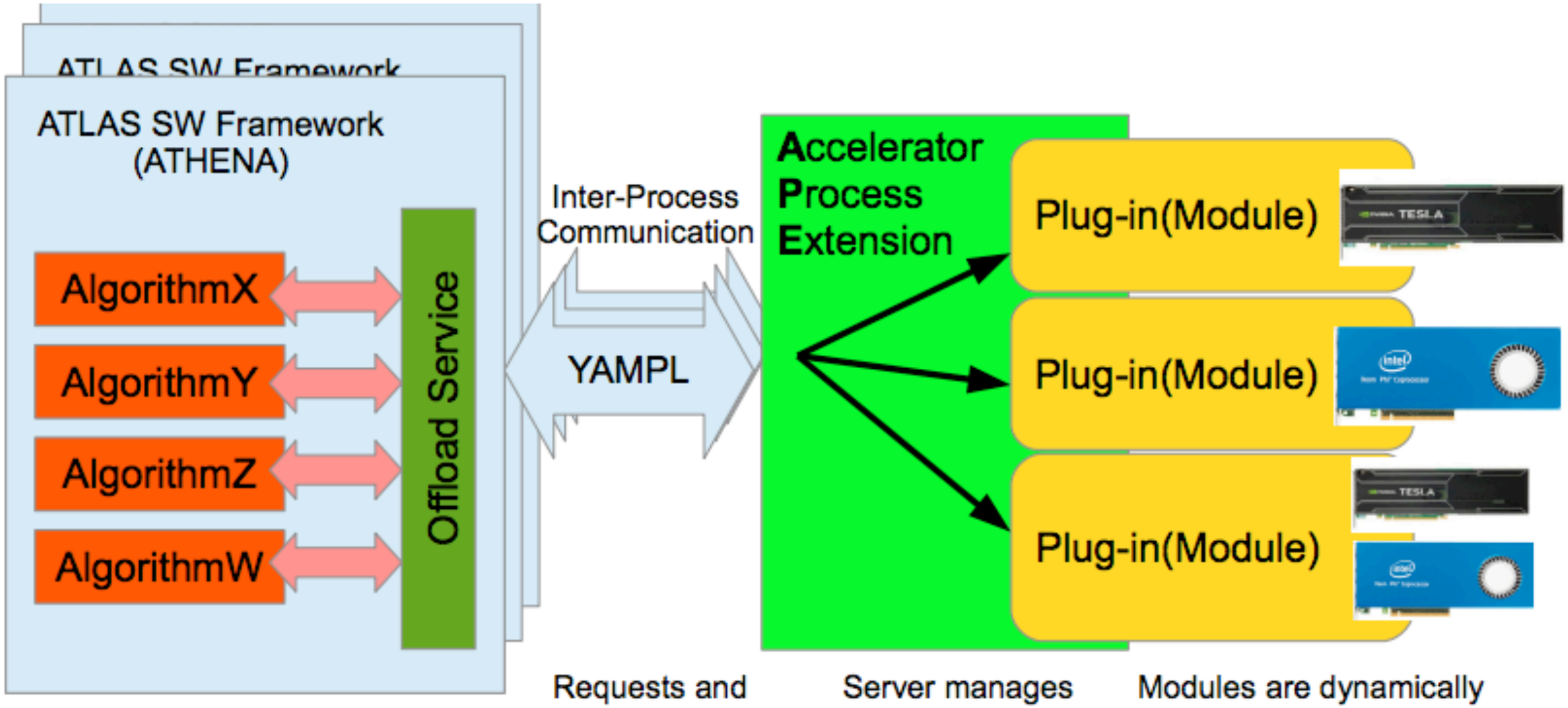
Trigger di alto e basso livello

- LHCb e’ abbastanza avanti nella ricostruzione di tracce nel trigger di alto livello (Hough Transform) su configurazioni manycore
    - Client-server tool per ‘offload’ di algoritmi da librerie
    - Varie esperienze di ricostruzione di tracce

- CMS
    - CUDA integrato in CMSSW ma anche indipendente: possibilita di lavorare con SW ibrido

- In ATLAS gli studi sono in uno stadio preliminare
    - Sviluppato un framework client-server per usare kernel plug-in nel framework ufficiale (Accelerator Process Extension)
        - possibile usare diversi linguaggi per diverse piattaforme
    - Studi preliminari frutto di iniziativa di singoli gruppi, al momento in fase di coordinamento

Trigger di alto e basso livello



- In ATLAS gli studi sono in uno stadio preliminare
  - Sviluppato un framework client-server per usare kernel plug-in nel framework ufficiale (Accelerator Process Extension)
    - possibile usare diversi linguaggi per diverse piattaforme
  - Studi preliminari frutto di iniziativa di singoli gruppi, al momento in fase di coordinamento

HEP: Lattice QCD, Monte Carlo, Event generation, Statisical Analysis

- LQCD: Approccio perturbativo non possibile comporta calcoli con matrici multidimensionali
    - Cruciale l'accesso memoria
    - Diverse suite (open source) su GPU sviluppate per i vari tipi di calcolo
- MC, EG
    - calcolo di sezione d’urto, generazione d’eventi, simulazione di apparati sperimentali

Astrofisica

- Large cosmological datasets per funzione di correlazione a 2 punti: problema assolutamente parallelizzabile da 100 a 300 di speedup

- Simulazioni dell'evoluzione delle strutture dell'universo: porting di codice "scafato" per GPU (OpenACC)

- Hermite Integration on GPU (HiGPU), codice per il problema a N corpi (integrazione numerica)

    - presi contatti per collaborazione per sperimentare sul cluster low-power

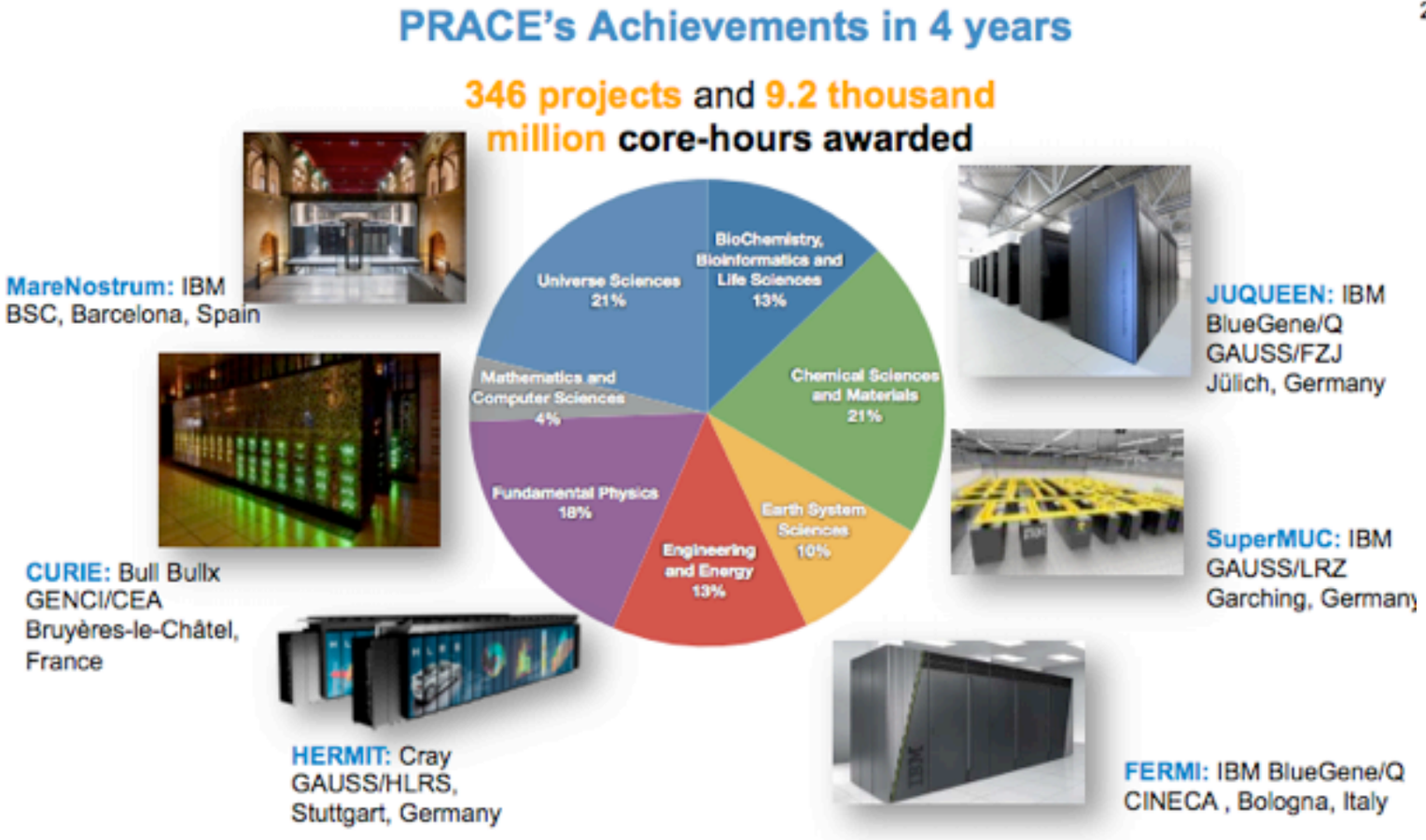
- Algoritmi con scala NlogN per problema a N corpi

Progetti/Consorzi

- GAP RT (GPU Application Project for Real Time)
    - accelerazione del calcolo scientifico con dispositivi commerciali: obiettivo di estensione al calcolo in tempo reale per applicazioni scientifiche e mediche (la sfida e' limitare la latenza aumentando la bandwidth e l'ottimizzazione in genere)
        - NMR, CT raggi X, PET
        - Trigger di basso livello per esperimenti (NA62, Km3)
- Montblanc: dispositivi embedded per hpc green, cheap, smart.
    - SoM: server on module! CPU+GPU+DRAM+storage+network in 8.5x5.6 cm
    - Eventuali problemi per messa in produzione hanno soluzioni budget-dependent

Progetti/Consorzi

- PRACE (Partnership for Advanced Computing in Europe): access to T0 and exascale systems



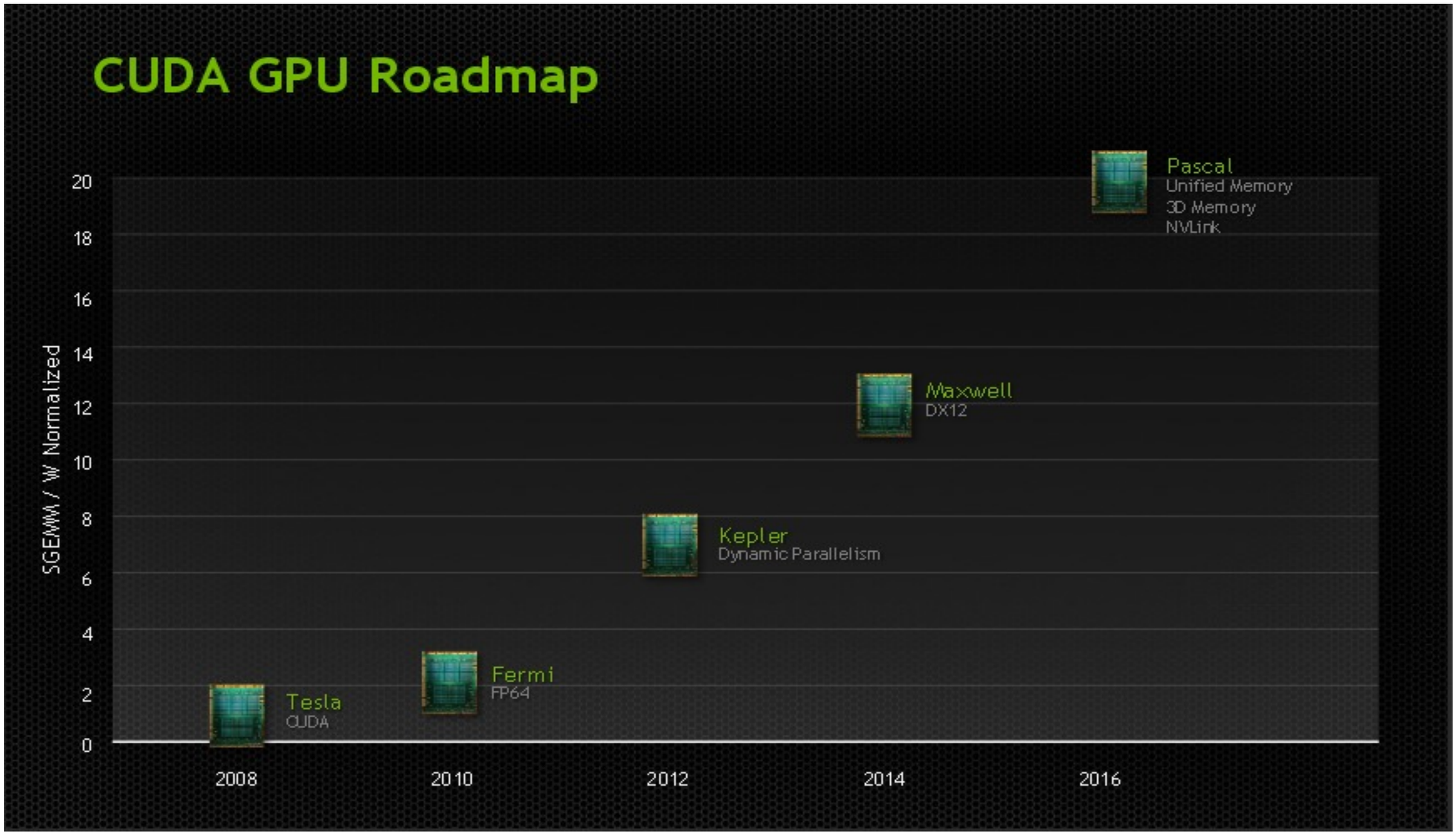
(2014)	Type A – 2 months	Type B/C – 6 months
Curie FN/TN Curie H	50.000 CPU 50.000 GPU	200.000 CPU 100.000 GPU
Hermit	50.000 CPU	50.000 CPU
FERMI	50.000 CPU	250.000 CPU
JUQUEEN	100.000 CPU	250.000 CPU
MareNostrum MareNostrum H	50.000 CPU 5.000 MIC	100.000 CPU 20.000 MIC
SuperMUC	100.000 CPU	250.000 CPU

Low power

- E4: benchmark su ARM64 + GPU non molto incoraggianti ne tantomeno low power

FEATURES	ARKA EK003 (2 server per chassis)
CPU	2x APM X-Gene A57 8 cores
GPU	2x NVIDIA Kepler® K2000 or K20
Memory	Up to 32 GB CPU RAM
Peak Performance	Up to 2.6 TFlops (1.3 TFlops per node)
Network	4x 10 GbE, 2x Infiniband FDR
Storage	2x SATA 2.0 Connector
Expansion slots	3x PCI-E 2.0 x4 (in x8), 1x PCI-E 2.0 x4 (in x16)

NVIDIA


CUDA GPU Roadmap
SGEMM / W Normalized
20
18
16
14
12
10
8
6
4
2
0
2008
2010
2012
2014
2016
Tesla
CUDA
Fermi
FP64
Kepler
Dynamic Parallelism
Maxwell
DX12
Pascal
Unified Memory
3D Memory
NVLink

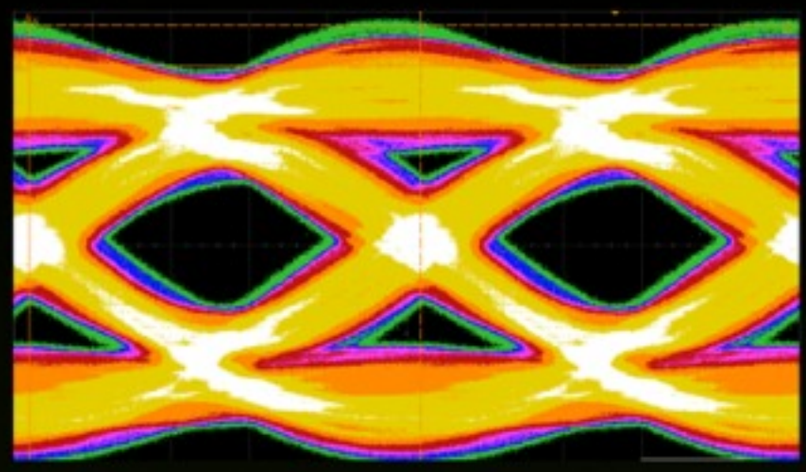
NVIDIA

NVLINK and Stacked Memory

NVIDIA

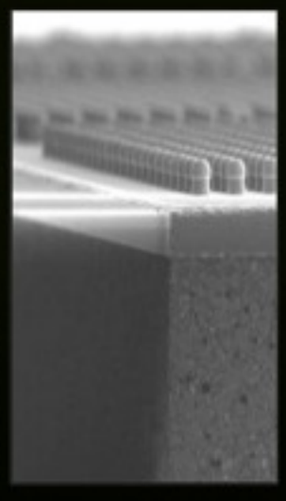
NVLINK

- GPU high speed interconnect
- 80-200 GB/s
- Planned support for POWER CPUs

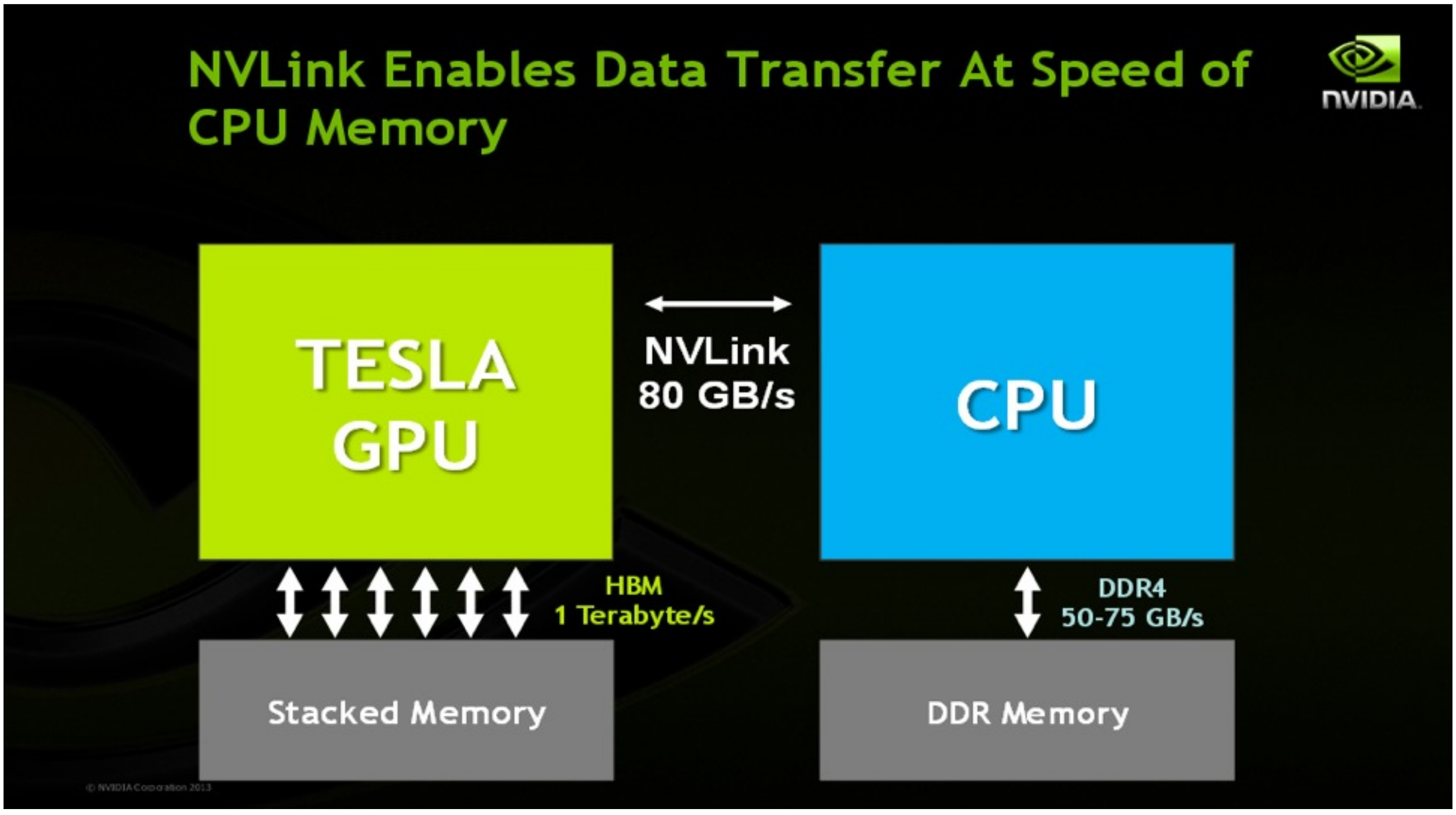


Stacked Memory

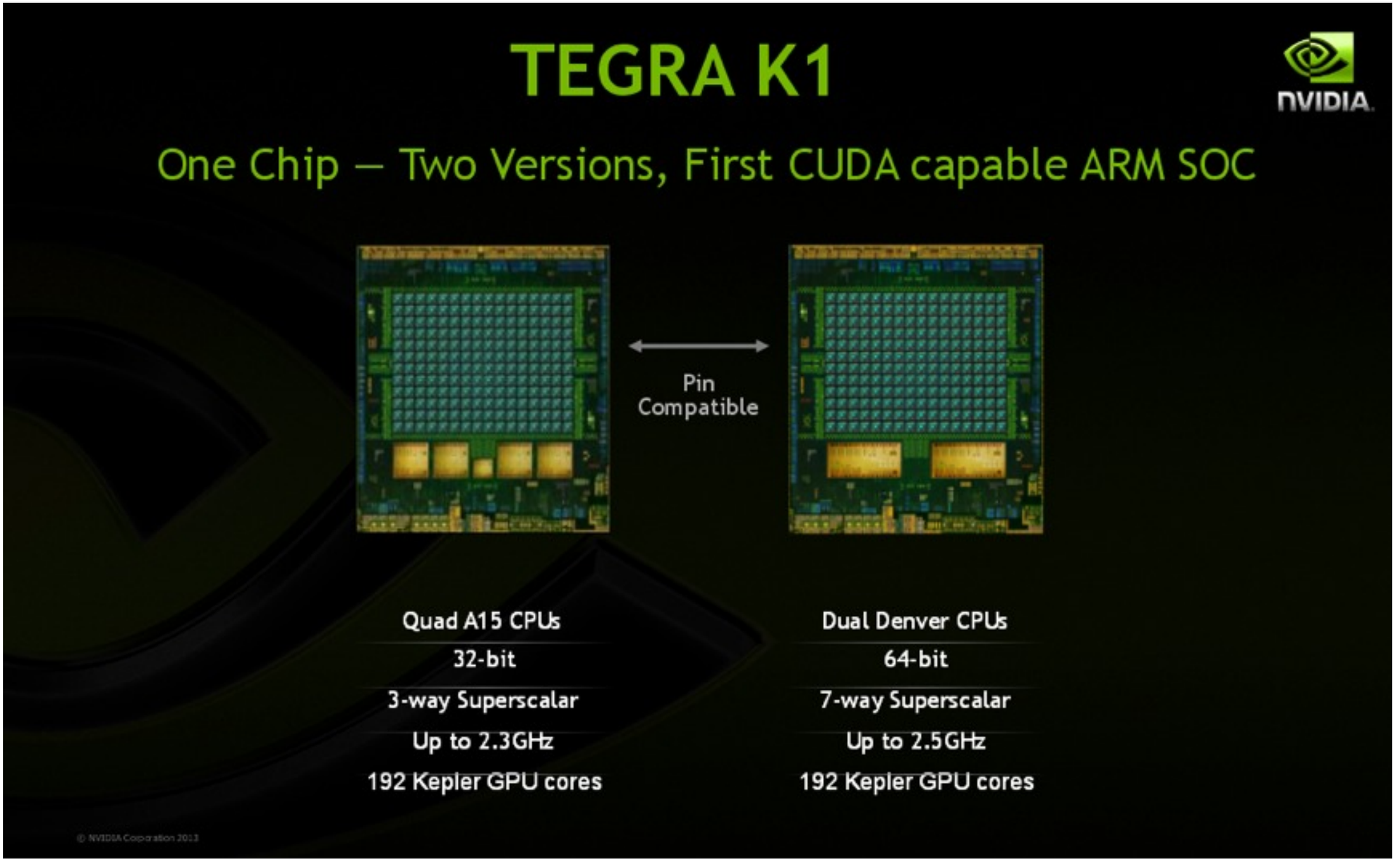
- 4x Higher Bandwidth (~1 TB/s)
- 3x Larger Capacity
- 4x More Energy Efficient per bit



NVIDIA


NVLink Enables Data Transfer At Speed of CPU Memory
NVIDIA
TESLA GPU
NVLink
80 GB/s
CPU
HBM
1 Terabyte/s
DDR4
50-75 GB/s
Stacked Memory
DDR Memory
© NVIDIA Corporation 2013

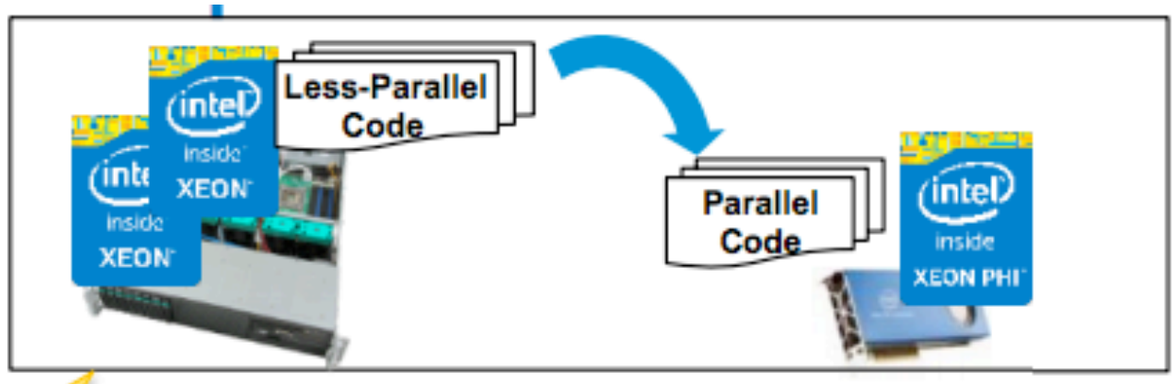
NVIDIA


TEGRA K1
NVIDIA
One Chip — Two Versions, First CUDA capable ARM SOC
Pin Compatible
Quad A15 CPUs
32-bit
3-way Superscalar
Up to 2.3GHz
192 Kepler GPU cores
Dual Denver CPUs
64-bit
7-way Superscalar
Up to 2.5GHz
192 Kepler GPU cores
© NVIDIA Corporation 2013

INTEL

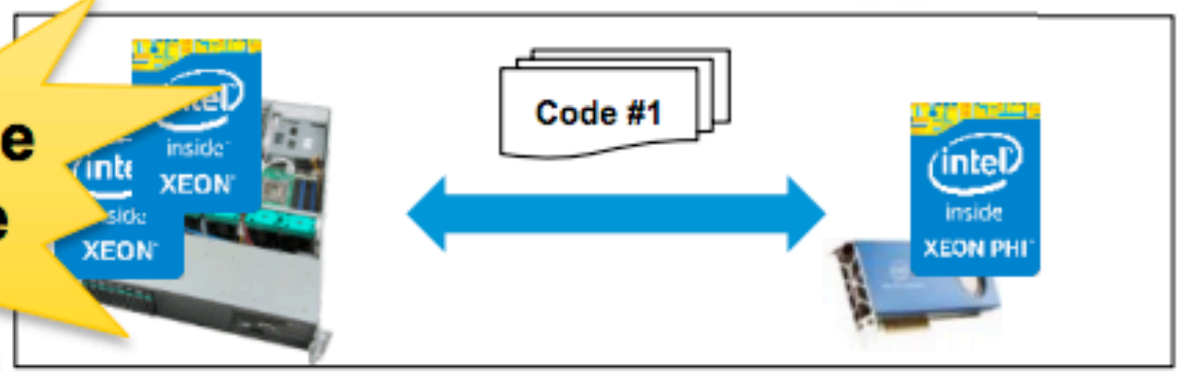
Splitting up a workload where highly-parallel code is offloaded to the coprocessor, and the Xeon® host processors primarily run primarily run less-parallel code

Nvidia GPUs only operate in OFFLOAD mode



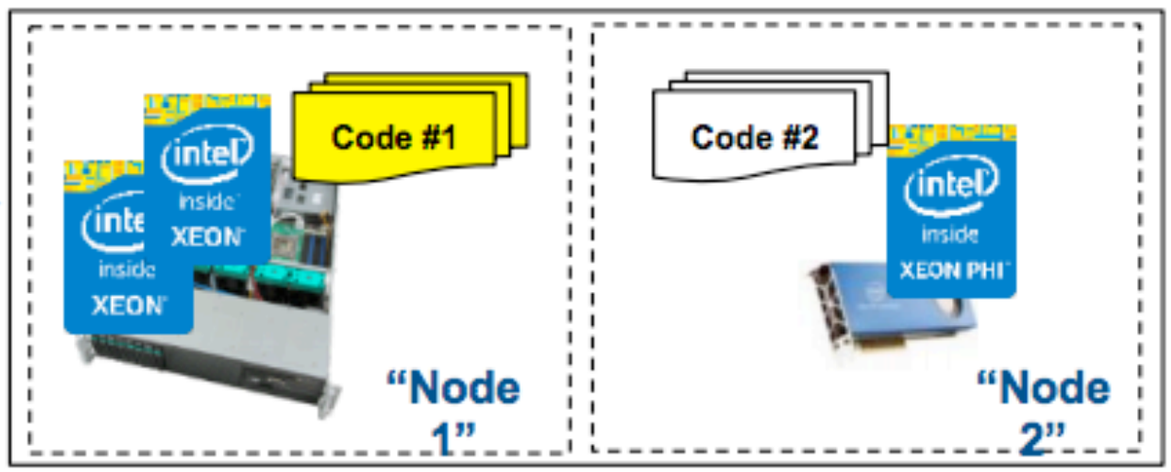
Workload is split up into multiple instances which are distributed and run on both the host processors and Intel® Xeon Phi™ coprocessors

Load Balance to Maximize Utilization



Coprocessor is configured as a separate compute node (full OS, IP addressable) to run an independent workload. Host CPUs are freed up to run separate workload(s)

Two Nodes, One System



INTEL

- Presentato il Knights Landing

- "Knights Landing" **code name** for the 2nd generation Intel® Xeon Phi™ product
- Based on Intel's **14 nanometer** manufacturing process
- Standalone **bootable processor** (running the host OS) and a **PCIe coprocessor** (PCIe end-point device)
- Integrated **on-package high-bandwidth memory**
- **Flexible memory modes** for the on package memory include: cache and flat
- Support for Intel® Advanced Vector Extensions 512 **(Intel® AVX-512)**
- ***60+ cores, 3+ TeraFLOPS*** *of double-precision peak performance per single socket node*
- ***Multiple hardware threads*** *per core with improved single-thread performance over the current generation Intel® Xeon Phi™ coprocessor*

- Punta sull'HPC portfolio, storage (Lustre), Fabric(StormLake), CPU, MIC

- How to get the best from the "XEON PHI" – veramente poco user friendly

conclusioni

- tanta ma tanta roba
  - molti esempi work in progress da cui poter trarre spunti o idee
  - progetti ambiziosi o interessanti
- a volte atteggiamenti un po troppo “dogmatici”
- tanti volenterosi e alle prime armi un po’ meno gli esperti consolidati e consapevoli