What comes after Yellowstone and how do we get ready for it?

Dr. Richard Loft
Director, Technology Development
Computational and Information Systems Laboratory
NCAR

CESM AMWG/WAWG
Feb 1, 2012

CISL Computational & Information Systems Laboratory

NCAR

Yellowstone

NWSC High Performance Computing Resource

- **Batch Computation Nodes**
  - 4,518 IBM dx360 M4 nodes; Intel Sandy Bridge EP† processors with AVX
  - 16 cores, 32 GB memory, 2.6 GHz clock, 333 GFLOPs per node
  - 4,518 nodes, 72,288 cores total – 1.504 PFLOPs peak
  - 144.6 TB total memory
  - 28.9 bluefire equivalents
- **High-Performance Interconnect**
  - Mellanox FDR InfiniBand full fat-tree
  - 13.6 GB/sec bidirectional bw/node
  - <2.5 usec latency (worst case)
  - 31.7 TB/sec bisection bandwidth
- **Login/Interactive Nodes**
  - 6 IBM x3650 M4 Nodes; Intel Sandy Bridge EP processors with AVX
  - 16 cores & 128 GB memory per node
- **Service Nodes (LSF, license servers)**
  - 6 IBM x3650 M4 Nodes; Intel Sandy Bridge EP processors with AVX
  - 16 cores & 32 GB memory per node

† "Sandy Bridge EP" is the Intel® Xeon® E5-2670 processor

CISL Computational & Information Systems Laboratory

NCAR

What comes after Yellowstone: Path to the Exascale (10^{18} flops)

System	Terascale (HPCx 2002)	Petascale (Jaguar 2009)	Exascale (DARPA strawman)
# of nodes	160	18,688	223,872
# cores/ node	8	12	742
# of cores	1280	224,256	166,113,024
# racks	40	284	583
Total Mem (TB)	1.28	300	3,580
Disk (TB)	18	600	3,580
Tape (TB)	35	10,000	3,580,000
Peak (Petaflop/s)	0.0067	2.33	1000
Total Power (MW)	0.5	7.0	68
Gflops/W	0.013	0.33	14.73
Bytes/Flop	0.5	0.2	0.0036

NCAR

Credit: Mike Ashworth, STFC, Daresbury, UK

Why are we turning to many-core?

- Since 2005 processor clock speeds have stagnated
- Why? Power consumption of high-GHz silicon
- Many-core design emphasizes executing many concurrent threads slowly, rather than executing a single thread very quickly.
- Where are we going? processors with hundreds of cores and thousands of threads

Why we're turning to many-core: Energy to do a double precision FLOP

- **Blue Fire (649 KW/59.7 TFLOPS)**
  - 10,873 pJ/FLOP
- **Yellowstone (1.9 MW/1.5 PFLOPS)**
  - 1,490 pJ/FLOP (huge improvement!)
- **Many-core systems:**
  - IBM Blue Gene/Q: 501 pJ/FLOP
  - NVIDIA KEPLER GPU: 200 pJ/FLOP (estimated)
  - Exascale target (DARPA): 68 pJ/FLOP = 68 MW system

Prototypes of next-gen architectures are coming and they will be disruptive

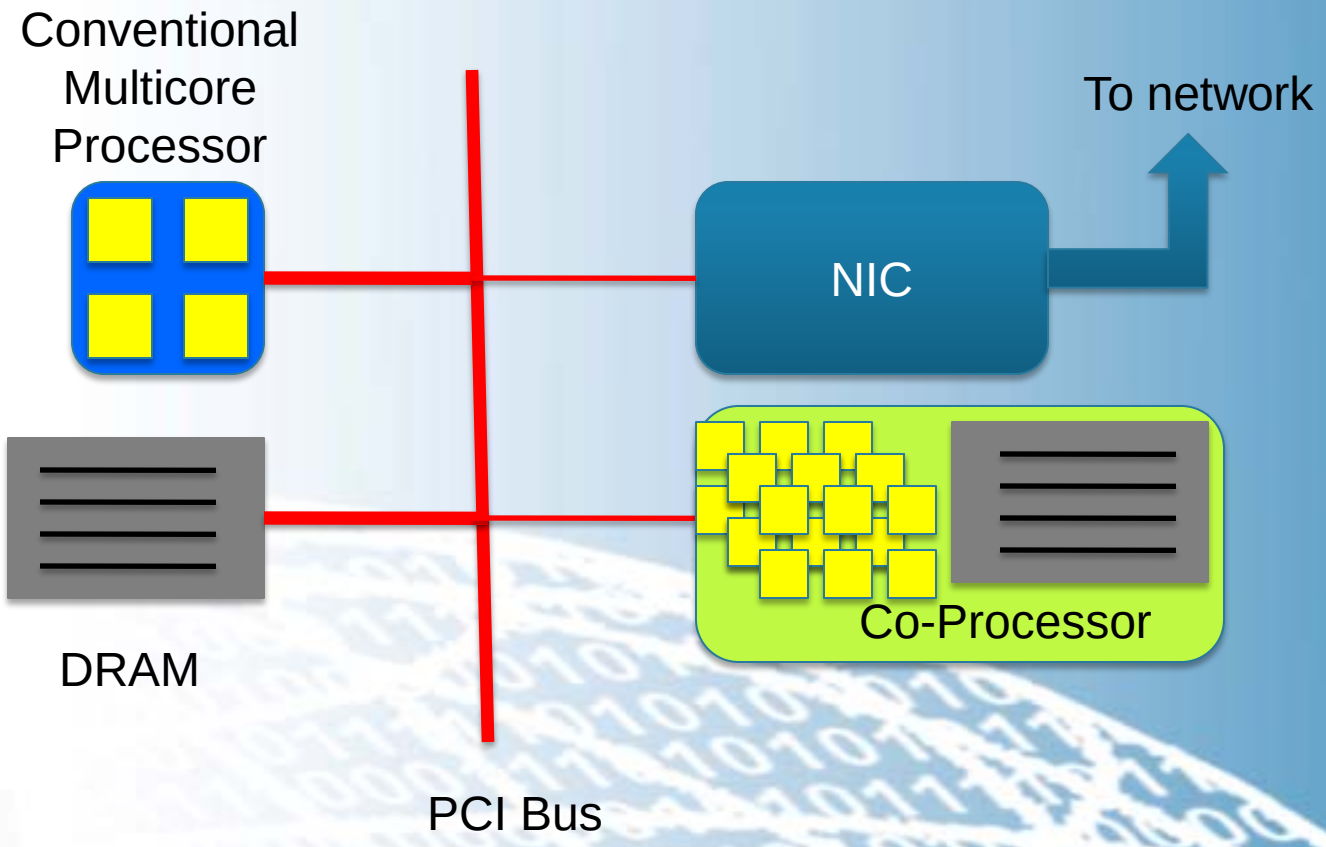
XSEDE: TACC Stampede (2013)
- 10 PF
- Intel MIC +Intel Sandy Bridge

DoE: ORNL Titan (2012)
- ~20 PF
- 18K NVidia Kepler GPU's + AMD Interlagos

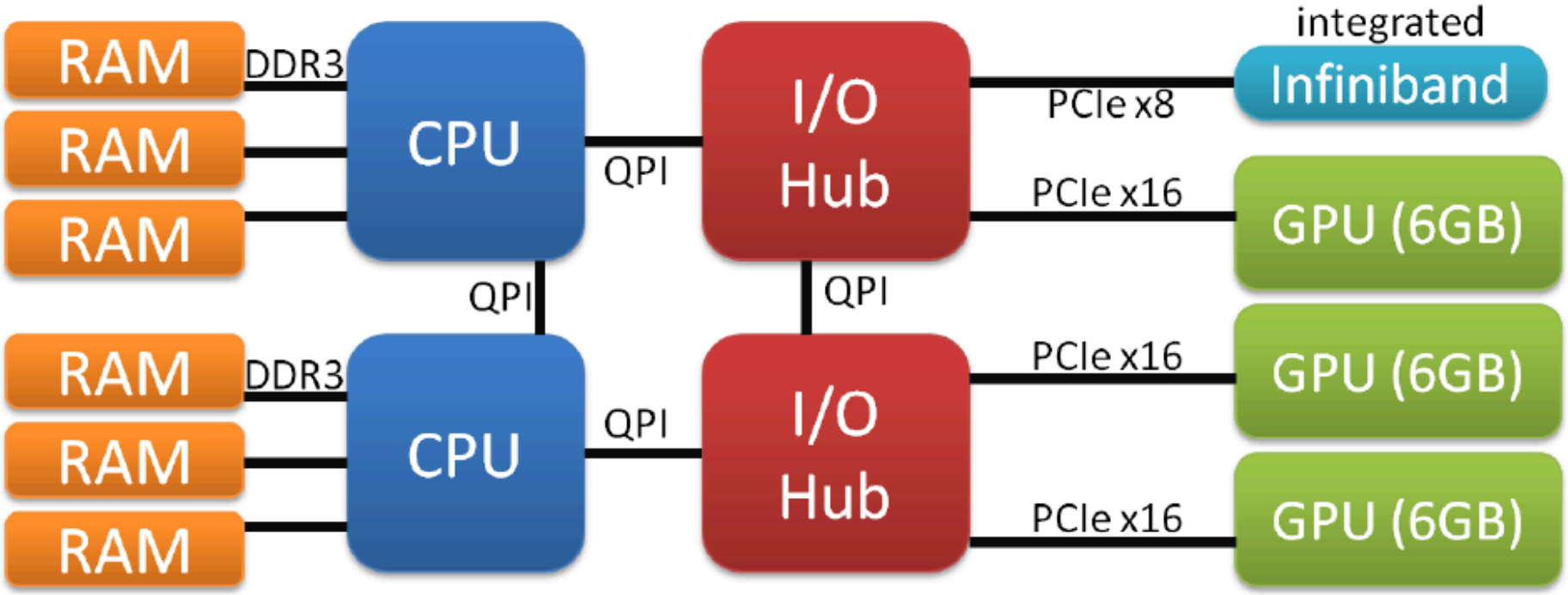
NSF Track 1: NCSA Blue Waters (2012)
- >11.5 PF
- AMD Interlagos + 3000 NVidia Kepler GPU's

DoE: Argonne Mira (2012)
- 9.2 PF
- 45,152 BG/Q System on a Chip (SoC)

Another Complication: The heterogeneous, co-processor node architecture



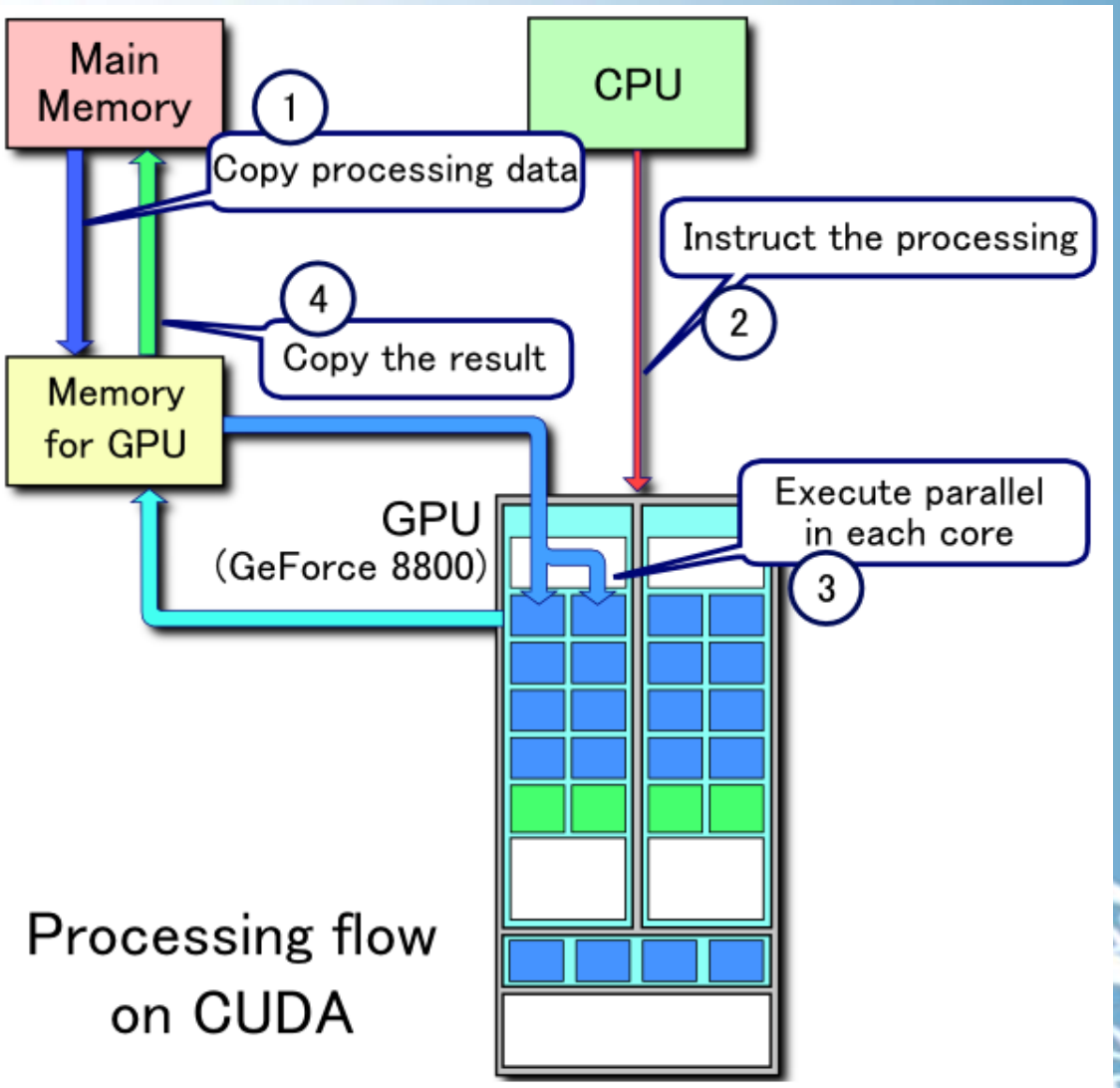
Example: “Keeneland” Node Architecture



NSF XSEDE Resource

Credit: Jeff Vetter, Georgia Tech

How to talk to coprocessors



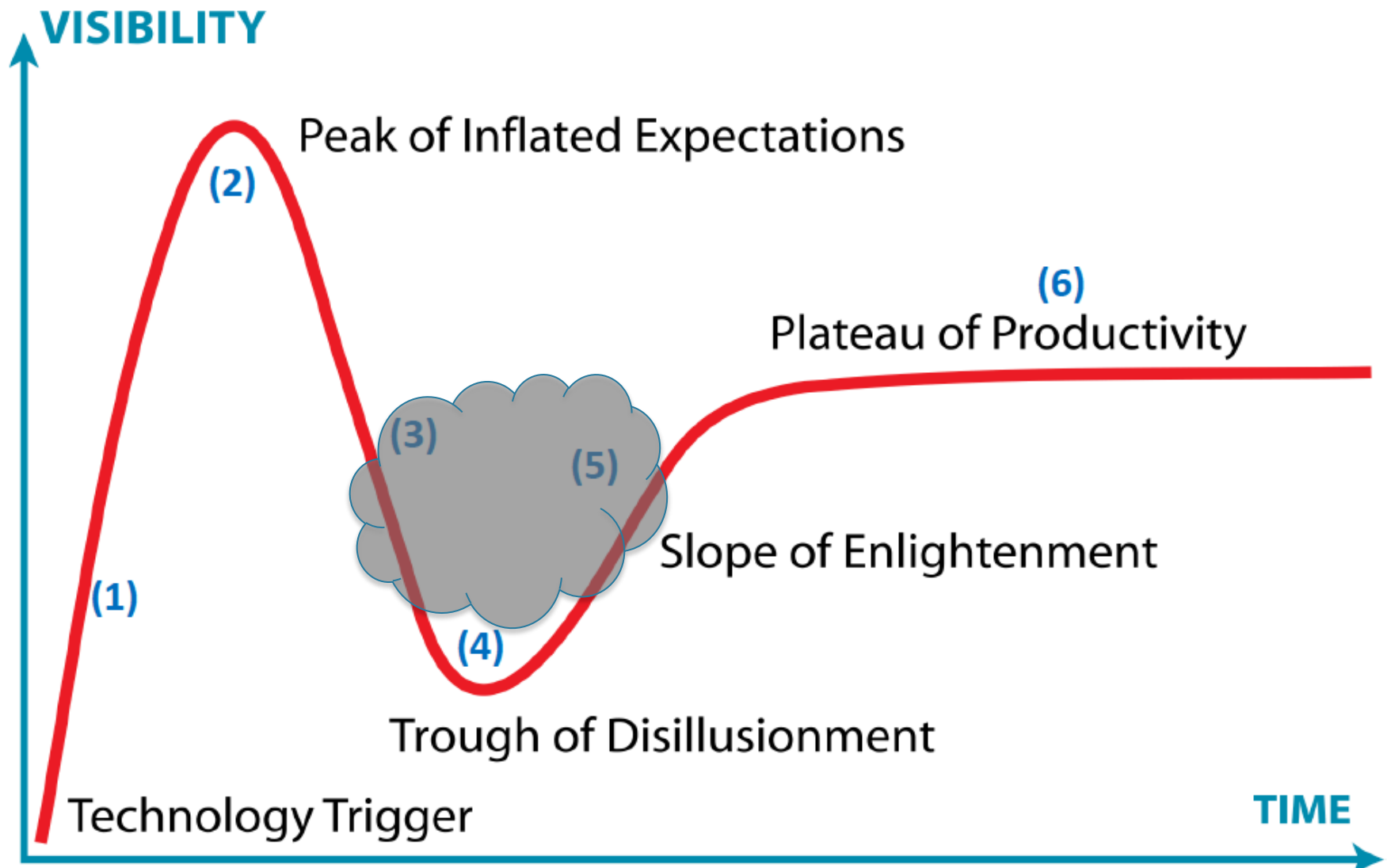
The Current Candidates…

BG/Q
Cores: 16
Multithread: 4-way
Coprocessor: no
Boot Linux: yes

Knights Ferry
Cores: 32
Multithread: 4-way
Coprocessor: yes
Boot Linux: yes

Fermi
Cores: 512
Multithread: 32-way
Coprocessor: yes
Boot Linux: no

Reality: where are we with many-core on the Gartner Hype cycle ?



Workshop to tackle the real issues: NCAR in September, 2011

Programming weather, climate, and earth-system models on heterogeneous multi-core platforms

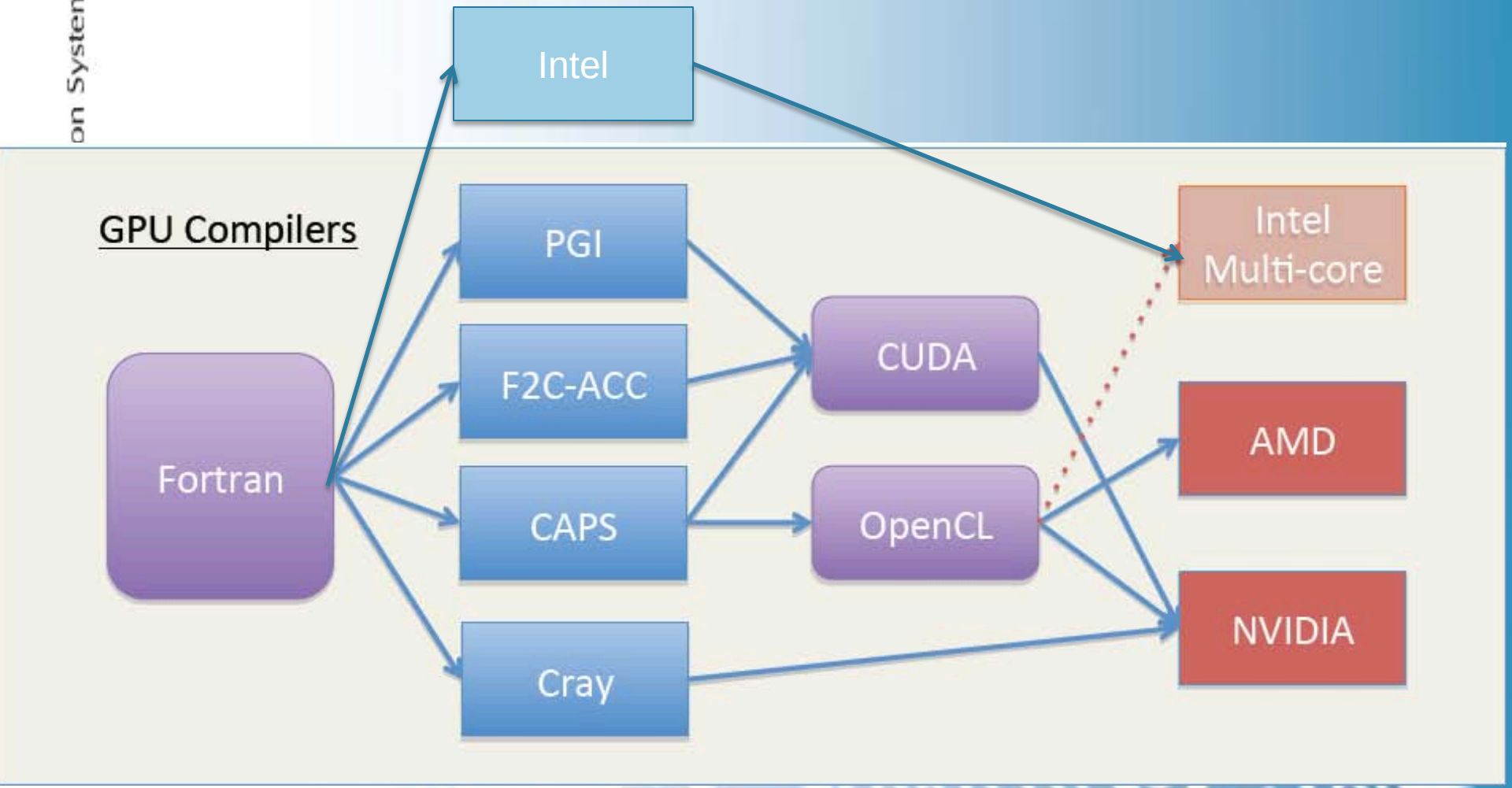
September 7-8, 2011 at the National Center for Atmospheric Research in Boulder, Colorado

GPU related talks (11+) that cover application software such as:

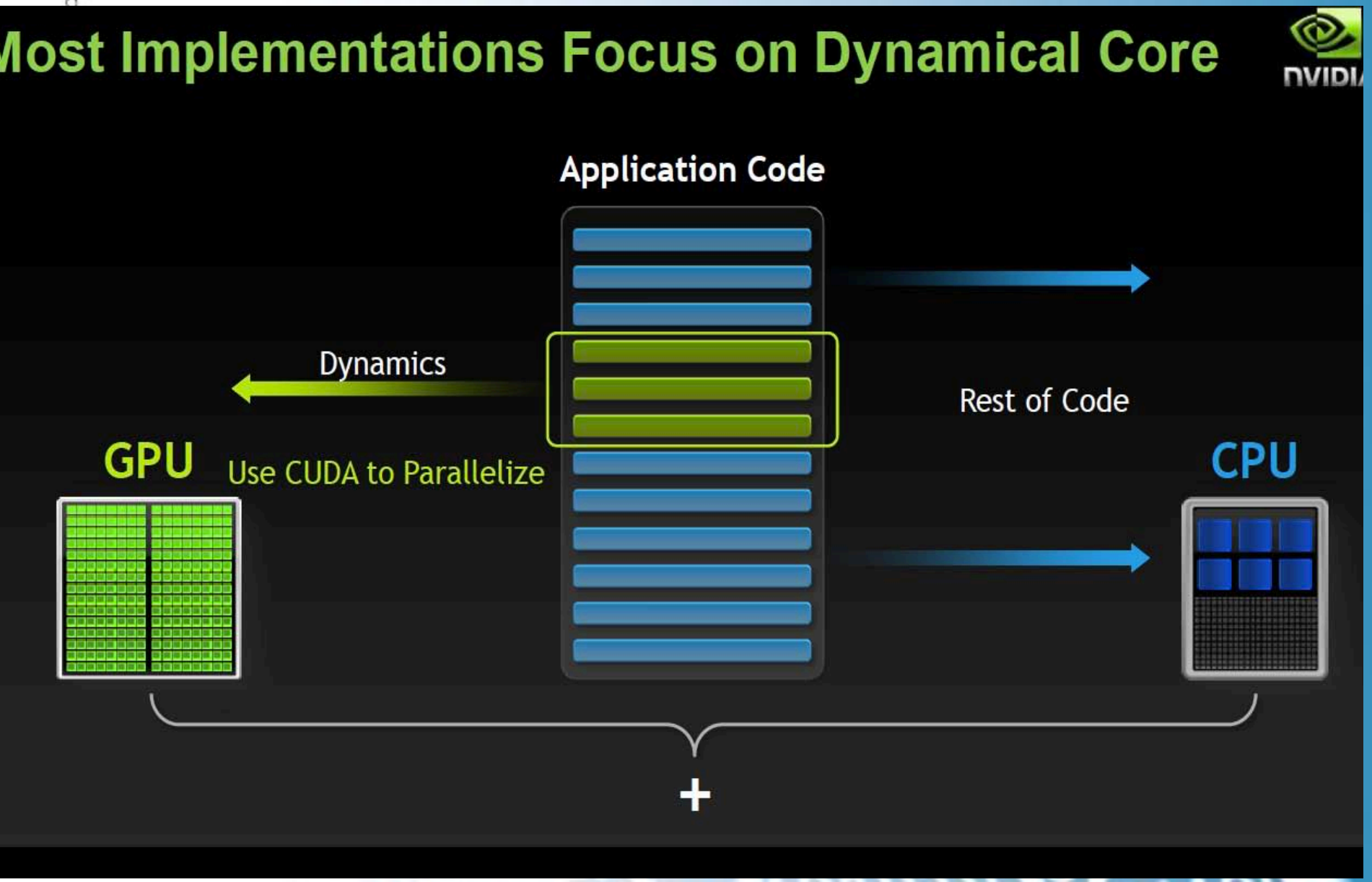
NIM | WRF | GEOS-5 | HOMME | COSMO | CAM | ICON

Successes and Challenges using GPUs for Weather and Climate Models	Mark Govett, NOAA
Experience using FORTRAN GPU Compilers with the NIM	Tom Henderson, NOAA
GPU Acceleration of the RRTM in WRF using CUDA FORTRAN	Greg Ruetsch, NVIDIA
Lessons Learned adapting GEOS-5 GCM Physics to CUDA FORTRAN	Matt Thompson, NASA
Accelerated Cloud Resolving Model in Hybrid CPU-GPU Clusters	Jose Garcia, NCAR
Reworking Boundary Exchanges in HOMME for Many-Core Nodes	Ilene Carpenter, NREL
Performance optimizations for running an NWP model on GPUs	Jacques Middlecoff, NOAA
Rewrite of the COSMO Dynamical Core	Mueller / Gysi, SCS/CSCS
Experiences with the Finite-Volume Dynamical core and GEOS-5 on GPUs	Bill Putman, NASA
Progress in Accelerating CAM-SE	Jeff Larkin, Cray/ORNL
Porting the ICON Non-hydrostatic Dynamical Solver to GPUs	Will Sawyer, CSCS

Current Pathways from Fortran to Hardware



First Porting Model: Dynamics First



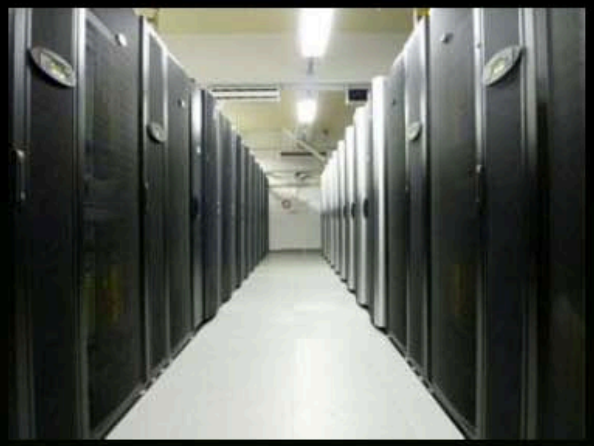
Second Model: Conventional Processor as Communication Co-Processor

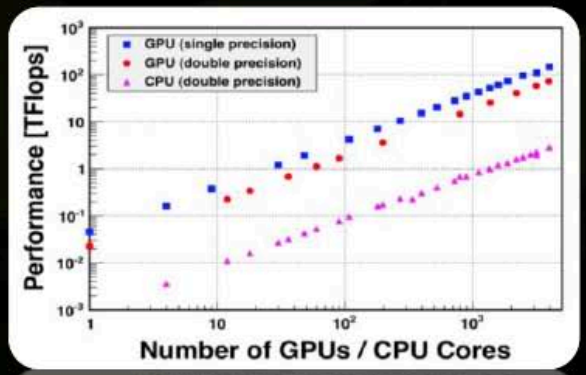
- **Invert traditional "GPU-as-co-processor" model**
  - Model state "lives" on GPU
  - Initial data read by the CPU and passed to the GPU
  - Data passed back to the CPU only for output & message-passing
  - GPU performs all computations
    - Fine-grained parallelism
  - CPU controls high level program flow
    - Coarse-grained parallelism
- **Minimizes overhead of data movement between CPU & GPU**

What can be achieved: ASUCA NWP on Tsubame 2.0

Tsubame 2.0
Tokyo Institute of Technology

- 1.19 Petaflops
- 4,224 Tesla M2050 GPUs





3990 Tesla M2050s
145.0 Tflops SP
76.1 Tflops DP

Simulation on Tsubame 2.0, TiTech Supercomputer

Concerns for CESM

- **Power crunch**
  - Utility/system fit-up costs for power-hungry systems are becoming cost prohibitive.
  - Many cores clocked @ 1.x GHz and thus use less power
- **Are our applications being left behind?**
  - If we can't use many-core systems effectively, what impact will that have on our science and on our programs?
  - Some work is being done in the community, but it appears "piecemeal" and under-resourced.
- **What's the right strategy?**
  - Slacker model – wait for SW/HW to "improve"…
  - Red-bull – hire a bunch of ace hackers and go for it?
  - Something in between?

We need an integrated assessment of CESM's many-core path forward:

CESM Science Objectives

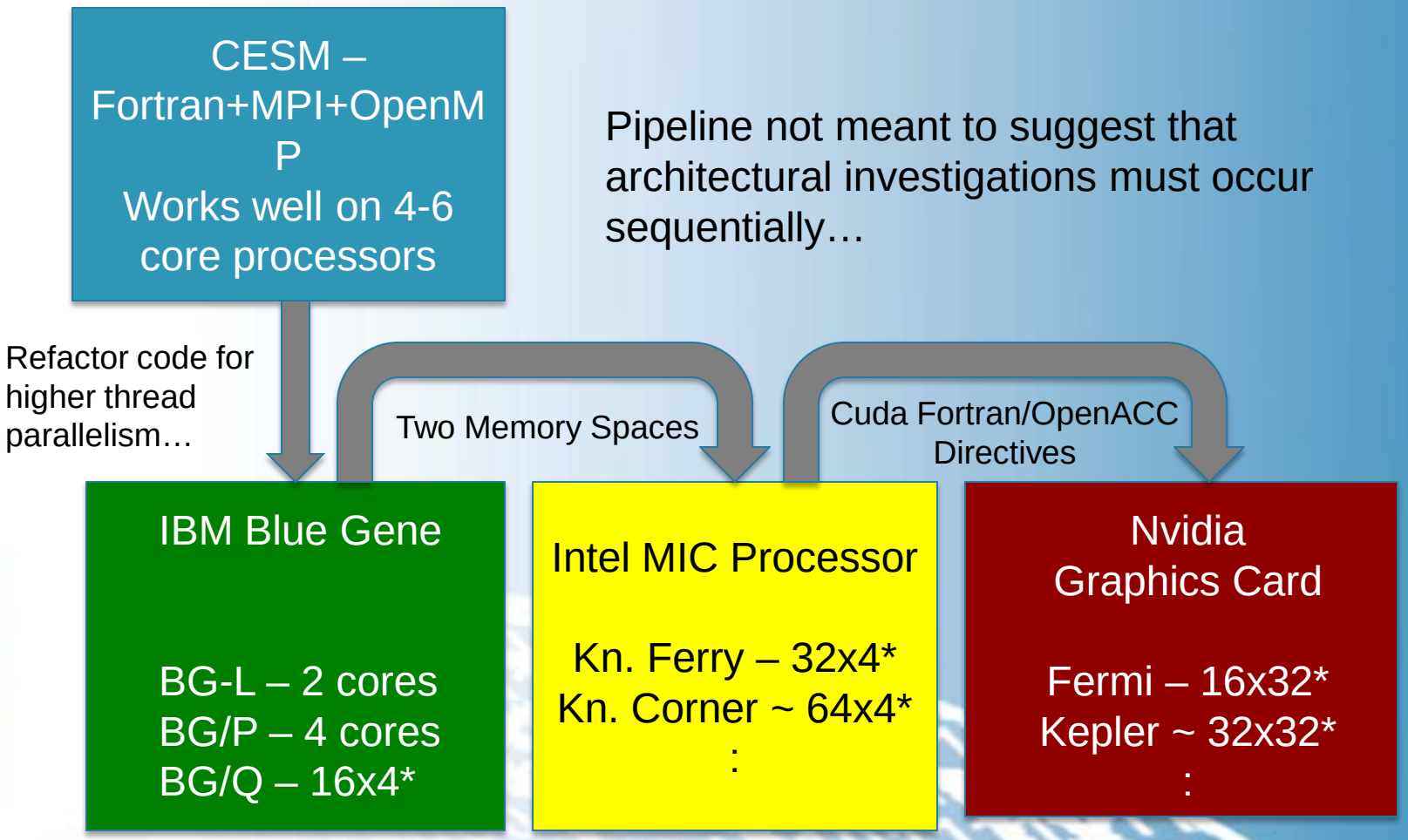
CESM Model Component Directions

Software Programming Models

Disruptive Technologies

CESM Disruptive Technology Working Group?

Possible Many-Core Path Forward



***cores X threads/core: most cores now have some form of multithreading**

This is hopefully the start of a broader discussion with the CESM community…

Thanks!

CISL Computational & Information Systems Laboratory

NCAR

Yellowstone Environment