Getting faster integration rates for paleoclimate simulations

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

CESM PCWG
Feb 17, 2012

CISL Computational & Information Systems Laboratory

NCAR

What I think I know about paleoclimate simulation

- **Paleoclimate (PC) runs at lower resolution for longer periods than typical runs of IPCC.**
  - O(10,000) year runs have been performed
  - 100 years/day vs 5 years/day
  - Resolutions coarser than 1 degree
- **Because of this, parallelism is a challeng.**
- **It generally is thought to runs best on fewer/faster processors.**

Some good news for paloeclimate on Yellowstone

- Yellowstone's Intel SB-EP processor cores **ought** to offer users about **50% more** throughput than bluefire POWER-6 cores.
- Vendor benchmark result: **CAM 0.25°** on 1000 cores gets **~3.0 years/day** on the previous generation of hardware.
- Working backward to a **1° PC**, that should reach **200 years/day** on **64** processors of Yellowstone.

Computer architectures are turning to massive numbers of slower threads. Why?

- 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

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

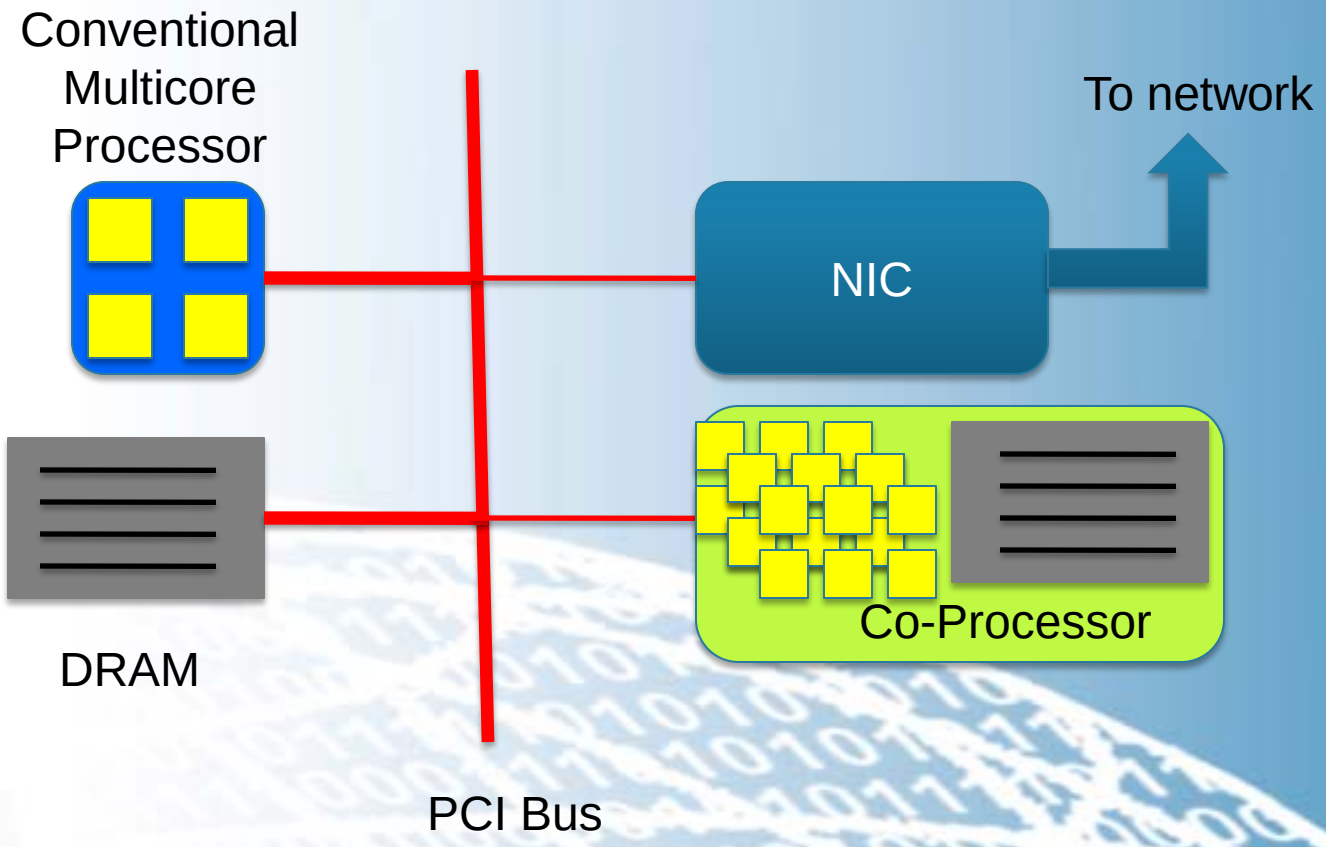
NCAR

Credit: Mike Ashworth, STFC, Daresbury, UK

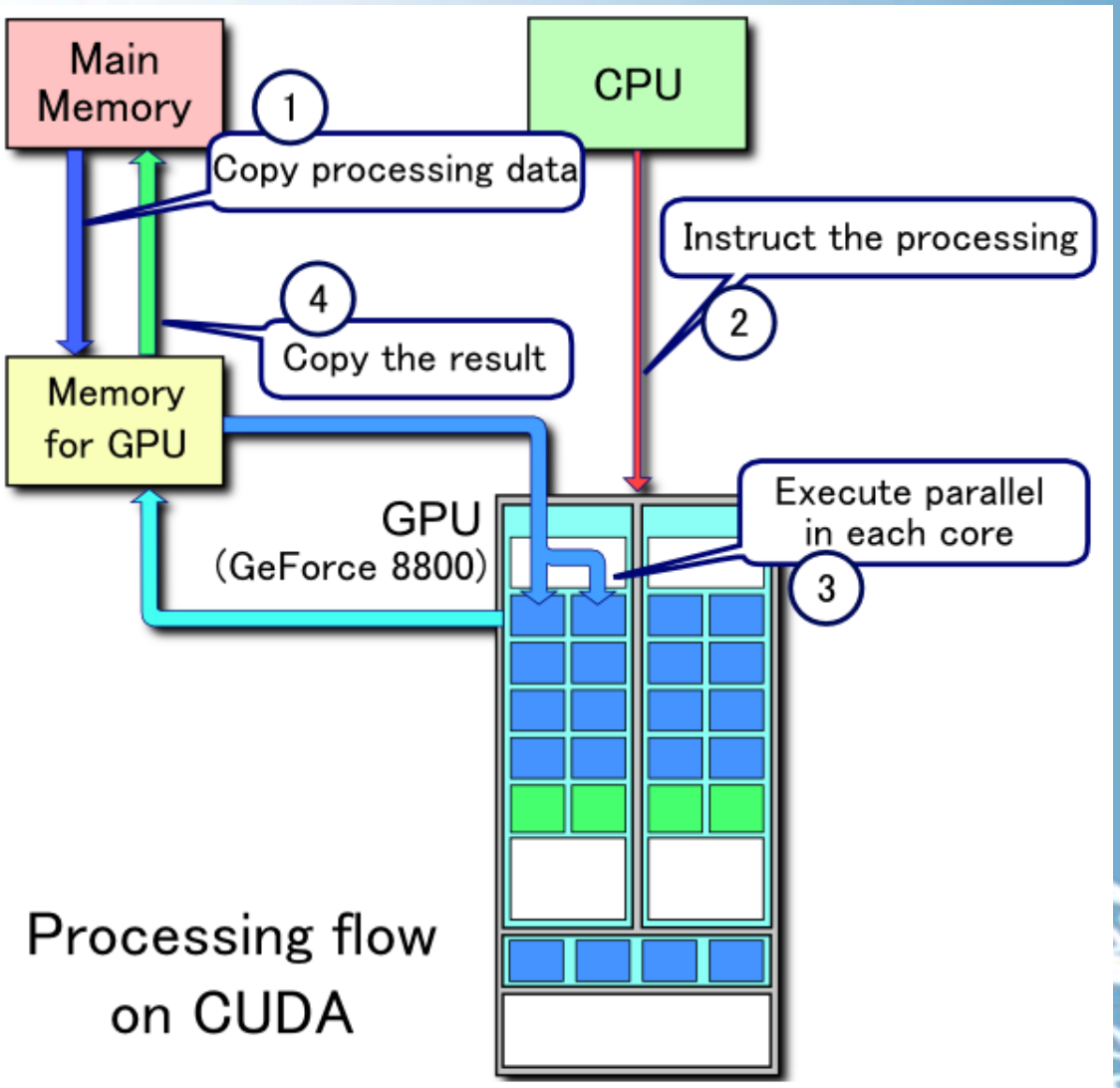
Why we're turning to many-core: Energy to do a double precision FLOP

- **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

Another Complication: The heterogeneous, co-processor node architecture

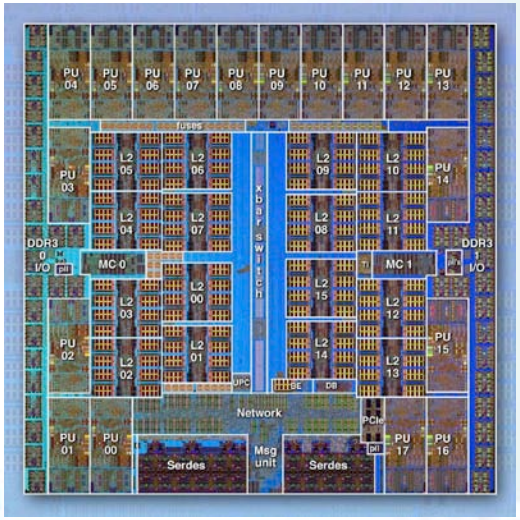


How to talk to coprocessors



Processing flow on CUDA

The Current Candidates…

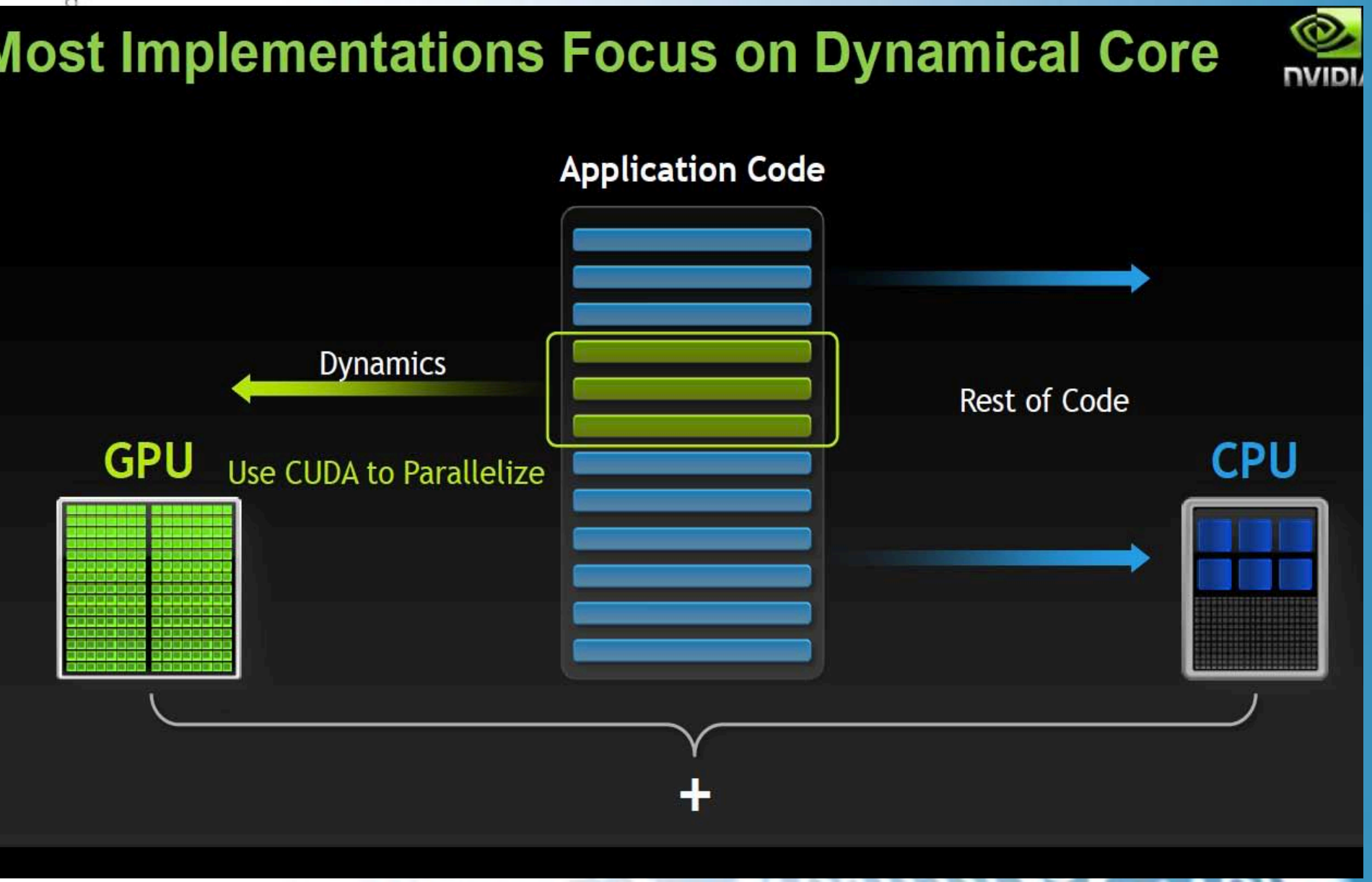


BG/Q
Cores: 16
Multithread: 4-way
Coprocessor: no
Boot Linux: yes

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

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

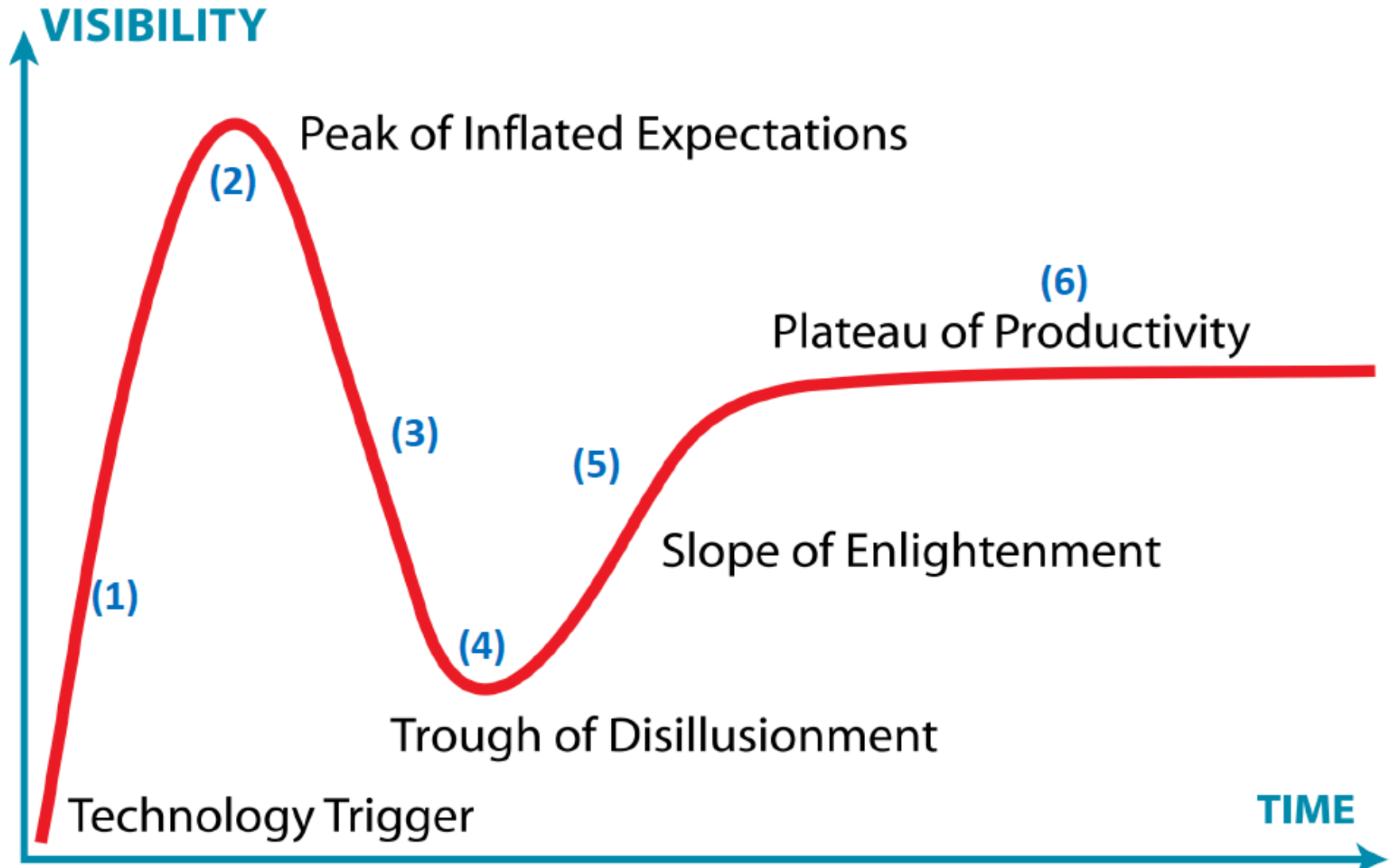
First Porting Model: Dynamics First



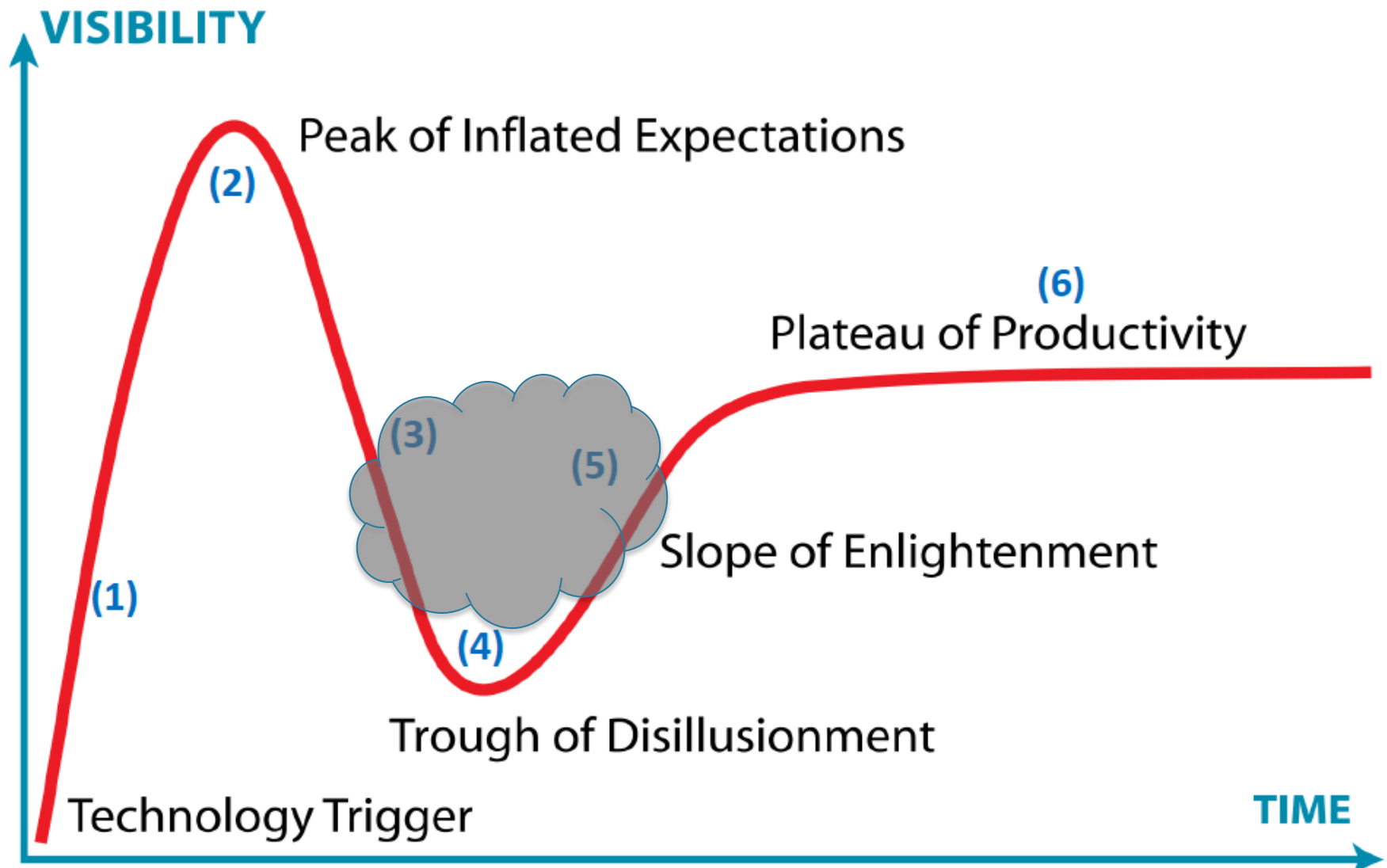
Second Model: Conventional Processor as Communication Co-Processor

- **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**

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



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



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?

Ideas for Special Purpose Paleoclimate Systems

- **Can Paleoclimate simulations ultimately fit on a single card?**
  - Forget MPI in this case, use threads
- **To go fast you need to maximize local memory bandwidth**
  - Graphics cards have very fast GDDR memory
  - Stacked (3D) memory in development for exascale

Stacked memory metaphor

Processor

Memory Stack

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

Thanks!

CISL Computational & Information Systems Laboratory

NCAR