Excerpts from
LA-UR-08-6246 &
LA-UR-08-2778 &
LA-UR-07-7405

Roadrunner:

A Fast But Unusual Bird, and Supercomputer

HPC User Forum
April 20, 2009

Ken Koch

Roadrunner Technical Manager,
Computer, Computational, and Statistical Sciences Division,
Los Alamos National Laboratory

Work presented was performed by a large team of Roadrunner project staff!

The messages this talk will convey are:

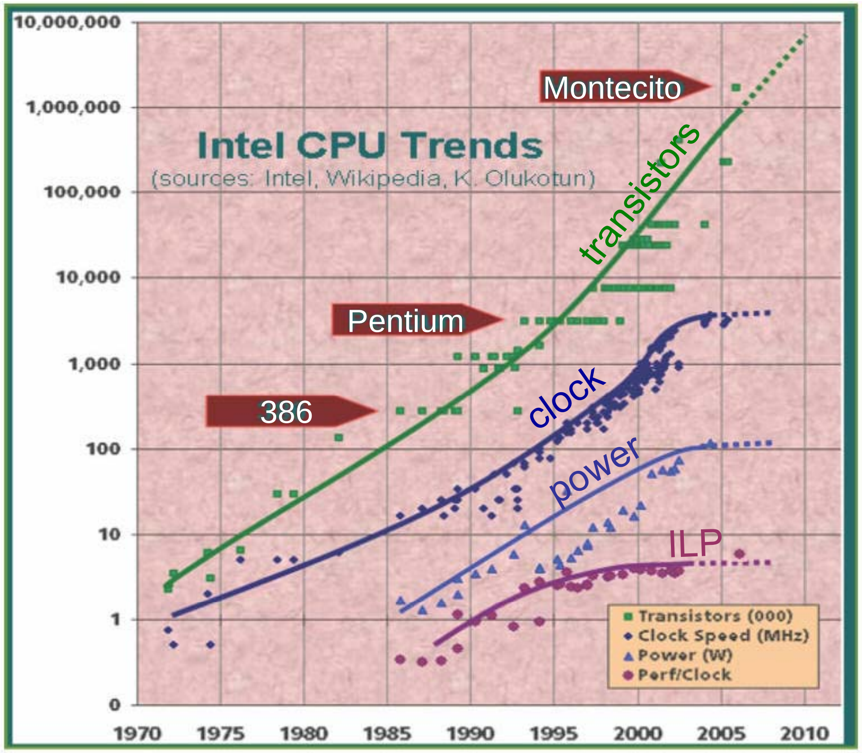
- Why Roadrunner? Why Cell?
  - *A bold but important step toward the future*
- What does Roadrunner look like?
  - *Cluster-of-clusters with node-attached Cell blades*
- Concepts for Programming Roadrunner
  - *MPI, Opteron+Cell, “local-store” memory & DMA transfers*
- Status and plans for Roadrunner
  - *Timeline*
  - *Applications to date*

A Roadrunner is born

using Cell processors as accelerators

Microprocessor trends have changed

- Moore’s law still holds, but is now being realized differently
  - *More cores per chip and not all cores need be the same*
  - *Decreased memory bandwidth and capacity per core*
  - *Key findings of Jan. 2007 IDC Study: “Next Phase in HPC”*
    - ***new ways of dealing with parallelism will be required***
    - *must focus more heavily on bandwidth (flow of data) and less on processor*

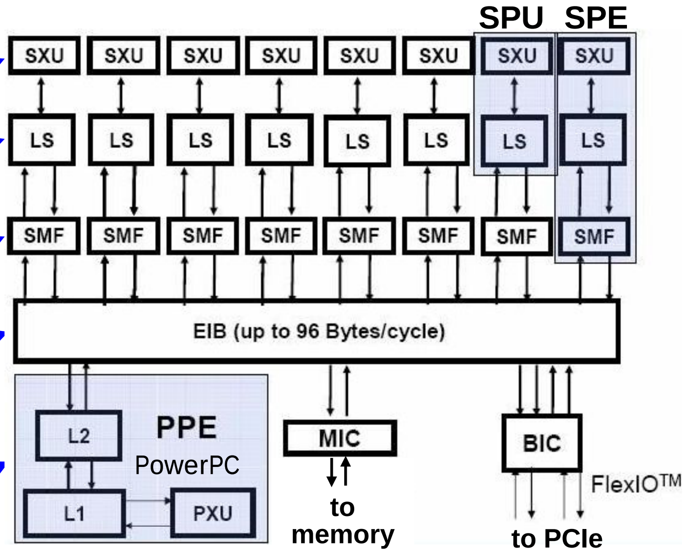


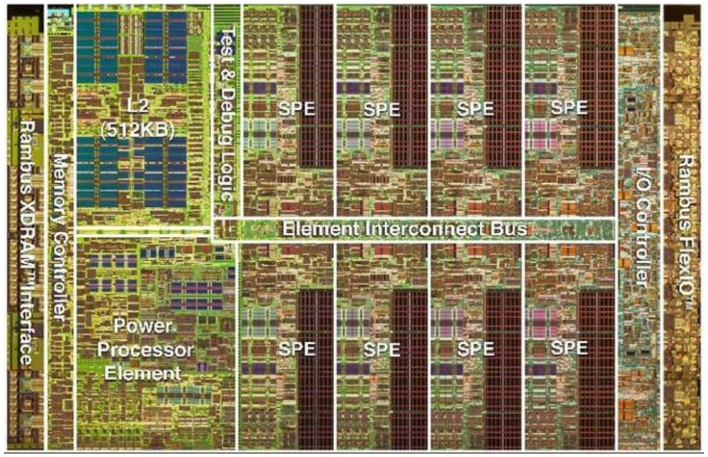
From Burton Smith, LASCI-06 keynote, with permission

The Cell processor is an (8+1)-way heterogeneous parallel processor

- Cell Broadband Engine (CBE*) developed by Sony-Toshiba-IBM
  - used in Sony PlayStation 3
- **8** Synergistic Processing Elements **(SPEs)**
  - 128-bit **vector cores**
  - 256 kB **local memory** (LS = Local Store)
  - Direct Memory Access **(DMA) engine** (25.6 GB/s each)
  - Chip interconnect (EIB)
  - Run SPE-code as POSIX threads (SPMD, MPMD, streaming)
- 1 PowerPC PPE runs Linux OS

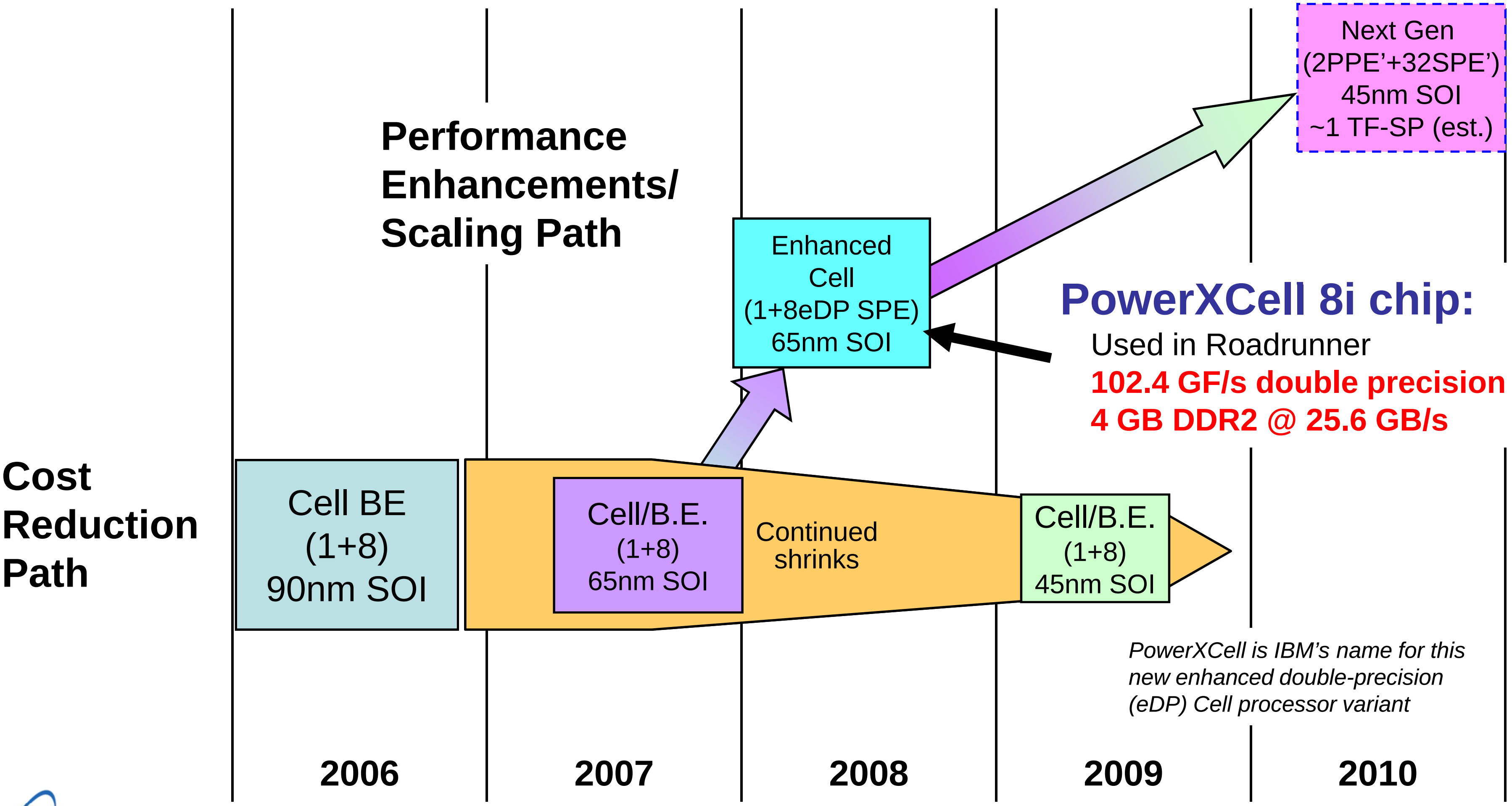
- Original Cell performance:
  - 204.8 GF/s SP & 13.65 GF/s DP
  - 512 MB @ 25.6 GB/s XDR memory
  - **Insufficient for a Petaflop/s machine for physics simulations**





* trademark of Sony Computer Entertainment, Inc.

IBM created PowerXCell 8i



PowerXCell 8i chip:

Used in Roadrunner

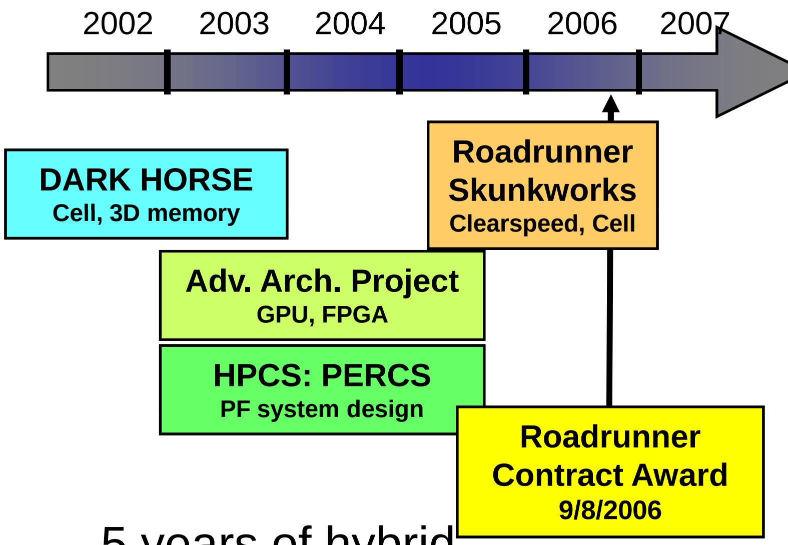
102.4 GF/s double precision

4 GB DDR2 @ 25.6 GB/s

PowerXCell is IBM's name for this new enhanced double-precision (eDP) Cell processor variant

All future dates and specifications are estimations only; Subject to change without notice. Dashed outlines indicate concept designs.

Los Alamos has a history in hybrid & petascale



DARK HORSE
Cell, 3D memory

Roadrunner Skunkworks
Clearspeed, Cell

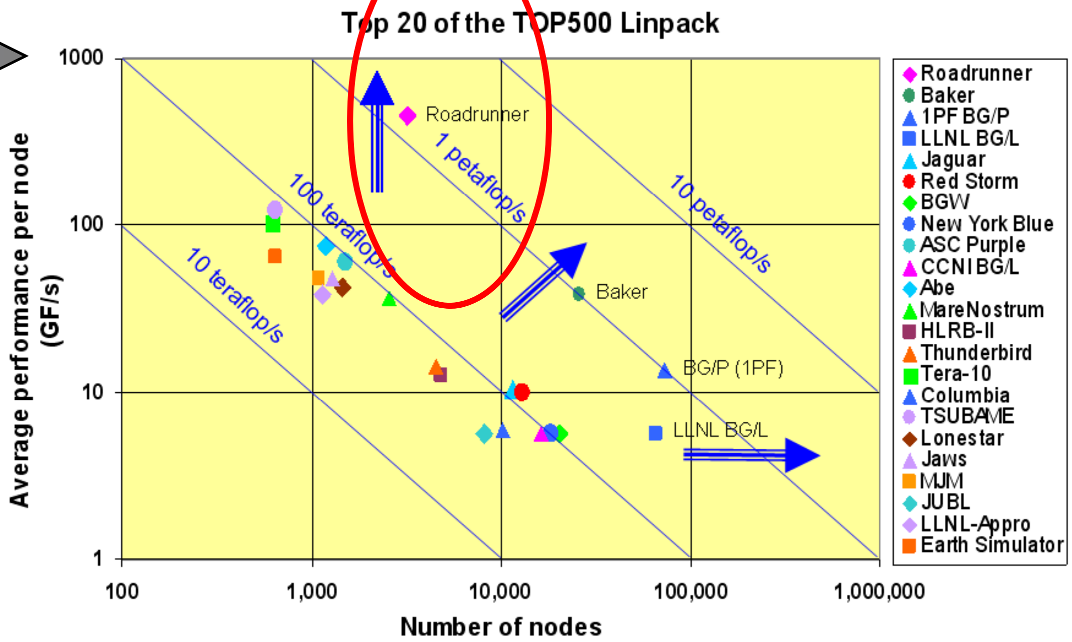
Adv. Arch. Project
GPU, FPGA

HPCS: PERCS
PF system design

Roadrunner Contract Award
9/8/2006

5 years of hybrid and petascale computing efforts led to Roadrunner

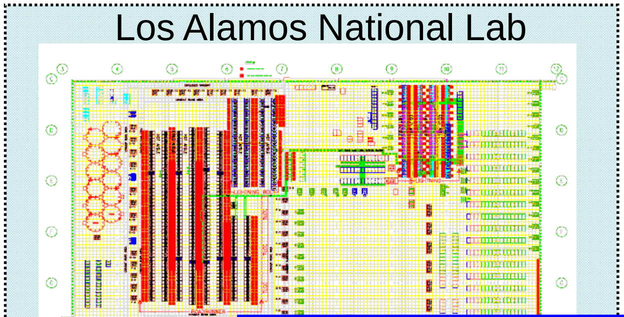
Paths to a Petaflop



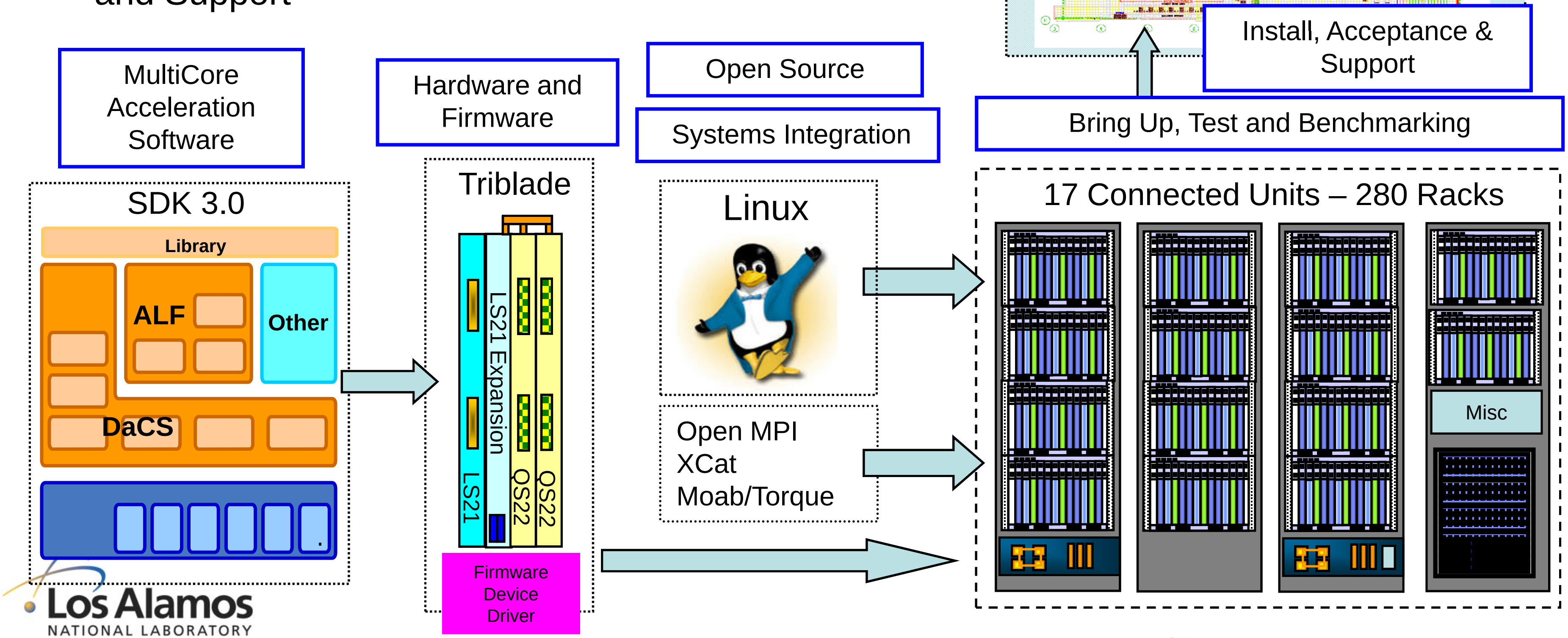
Roadrunner took a different path to a petascale system

Roadrunner was delivered to LANL in Summer 2008

- New Hybrid Computing Model & Software
- New Triblade Hybrid Compute Node
- IBM Global Engineering Solutions was responsible for Development, Manufacturing and Support



Los Alamos National Lab



Install, Acceptance & Support

MultiCore Acceleration Software

Hardware and Firmware

Open Source

Systems Integration

Bring Up, Test and Benchmarking

SDK 3.0

Triblade

Firmware Device Driver

Linux

Open MPI
XCat
Moab/Torque

17 Connected Units – 280 Racks

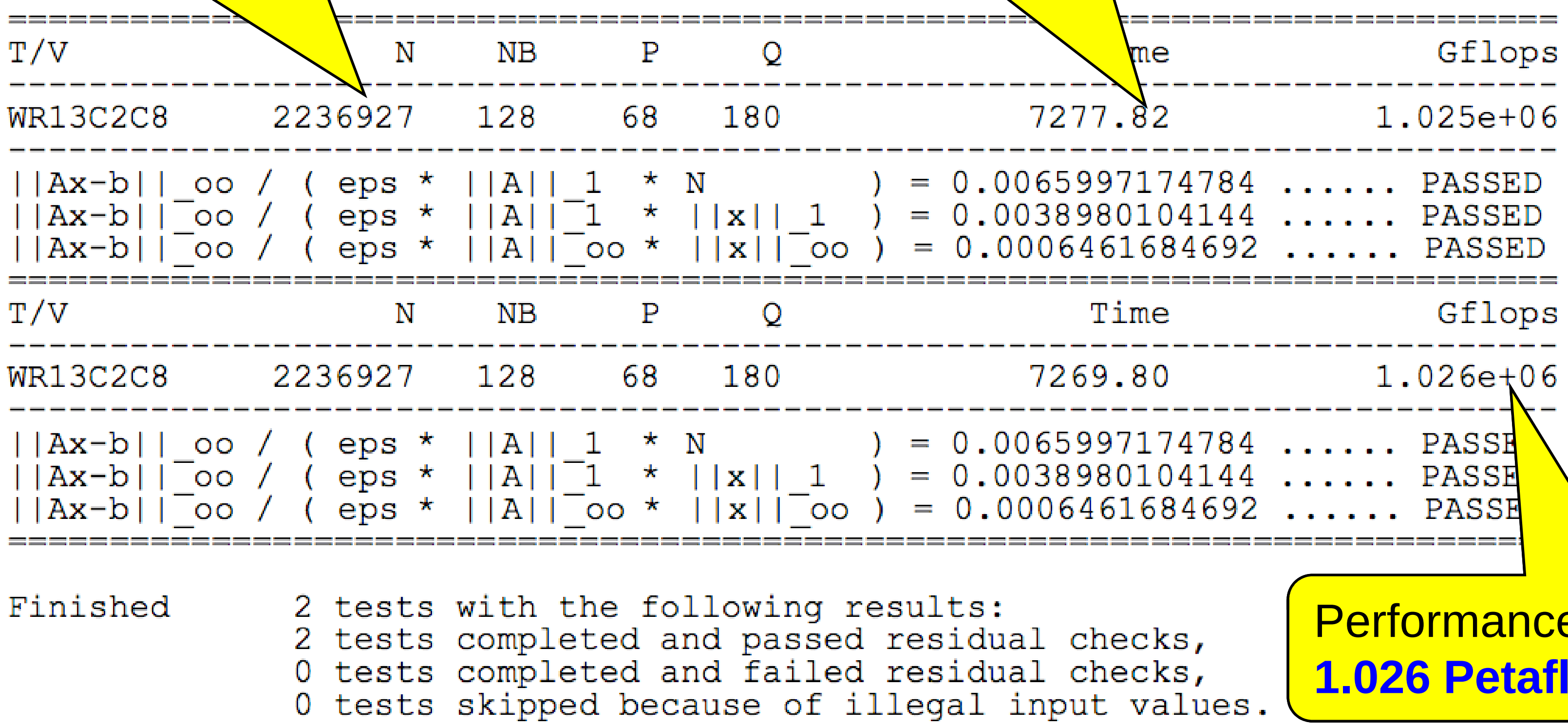
IBM built hybrid nodes in Rochester, MN and assembled the system in Poughkeepsie, NY

Fully Assembled Roadrunner

Roadrunner broke the 1 Petaflop/s mark on May 26th, 2008

Matrix: ~5 trillion entries

Calculation: **~2 hours!**



```
================================================================================
T/V                N    NB     P     Q                me             Gflops
--------------------------------------------------------------------------------
WR13C2C8     2236927   128    68   180             7277.82          1.025e+06
--------------------------------------------------------------------------------
||Ax-b||_oo / ( eps * ||A||_1  * N        ) = 0.0065997174784 ...... PASSED
||Ax-b||_oo / ( eps * ||A||_1  * ||x||_1  ) = 0.0038980104144 ...... PASSED
||Ax-b||_oo / ( eps * ||A||_oo * ||x||_oo ) = 0.0006461684692 ...... PASSED
================================================================================
T/V                N    NB     P     Q              Time             Gflops
--------------------------------------------------------------------------------
WR13C2C8     2236927   128    68   180             7269.80          1.026e+06
--------------------------------------------------------------------------------
||Ax-b||_oo / ( eps * ||A||_1  * N        ) = 0.0065997174784 ...... PASSE
||Ax-b||_oo / ( eps * ||A||_1  * ||x||_1  ) = 0.0038980104144 ...... PASSE
||Ax-b||_oo / ( eps * ||A||_oo * ||x||_oo ) = 0.0006461684692 ...... PASSE
================================================================================

Finished      2 tests with the following results:
              2 tests completed and passed residual checks,
              0 tests completed and failed residual checks,
              0 tests skipped because of illegal input values.
```

Performance:
1.026 Petaflop/s

Only 4 days after the full machine was finally assembled!

Roadrunner is a TOP performer!

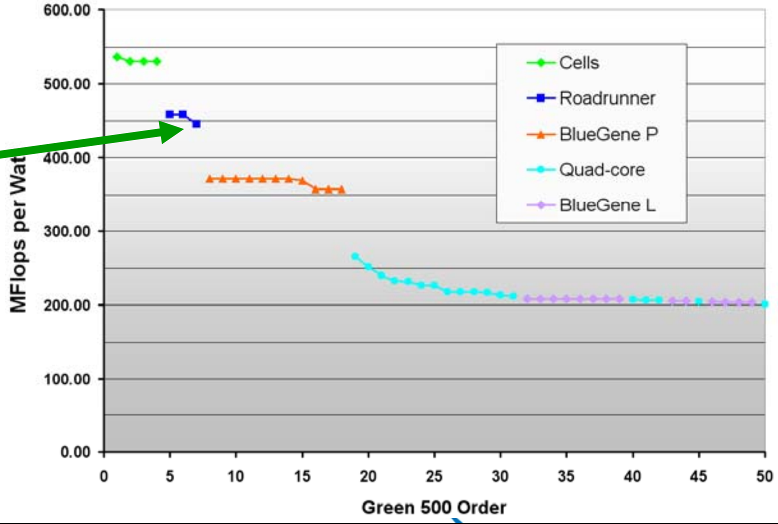
#	SITE	SYSTEM	Cores	TF/sec	MW
1	**DOE/NNSA/LANL** **United States**	**Roadrunner, QS22/LS21/IB** **PowerXCell 8i, IBM**	**129600***	**1105**	**2.48**
2	DOE/ORNL United States	Jaguar, XT5, Opteron-QC, Cray	150152	1059	6.95
3	NASA Ames Research Center United States	Pleiades, Altix ICE & IB, Xeon-QC SGI	51200	487	2.09
4	DOE/NNSA/LLNL United States	BGL, Blue Gene/L, PowerPC, IBM	212992	478	2.33
5	Argonne National Laboratory United States	Intrepid, Blue Gene/P, PowerPC, IBM	163840	450	1.26
6	Texas Adv. Comp. Center United States	Ranger, SunBlade & IB Opteron-QC, Sun	62976	433	2.00

#1 on the TOP500
(Nov. 2008)

* Roadrunner core count
includes 8 SPEs in each Cell;
Opteron+Cell-PPE cores is
only 25920

#7 on the Green500
(Nov. 2008)

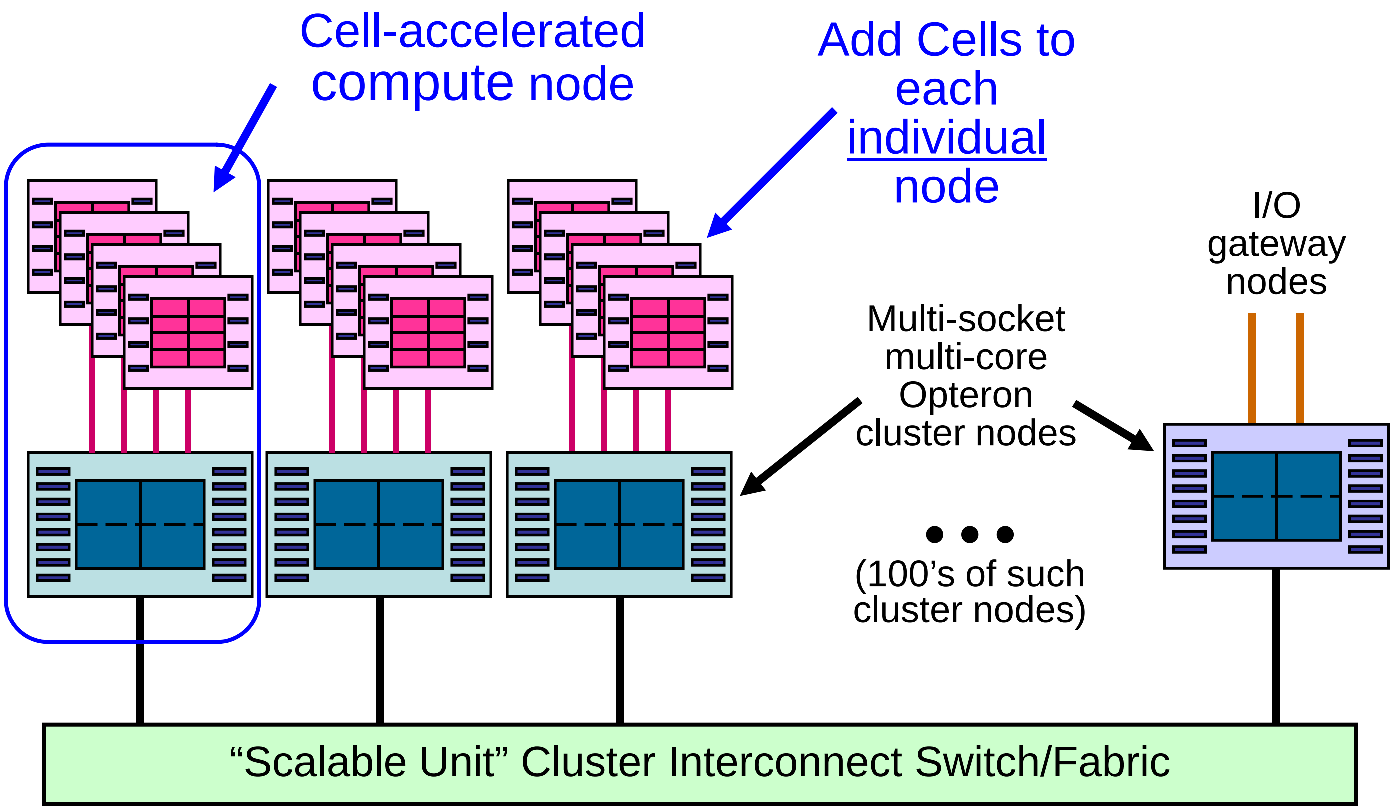
#5 & #6 are Roadrunner
single CUs at LANL & IBM



Roadrunner System Configuration

See the LANL Roadrunner web site at end for more details

Roadrunner Phase 3 is Cell-accelerated, not a cluster of Cells



Cell-accelerated compute node

Add Cells to each individual node

Multi-socket multi-core Opteron cluster nodes

I/O gateway nodes

• • •

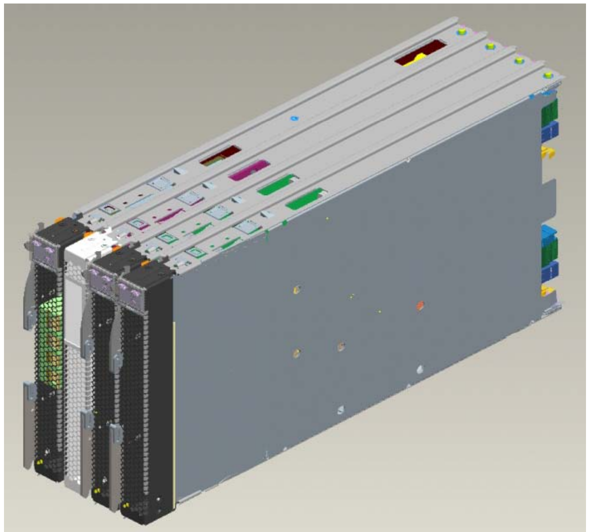
(100’s of such cluster nodes)

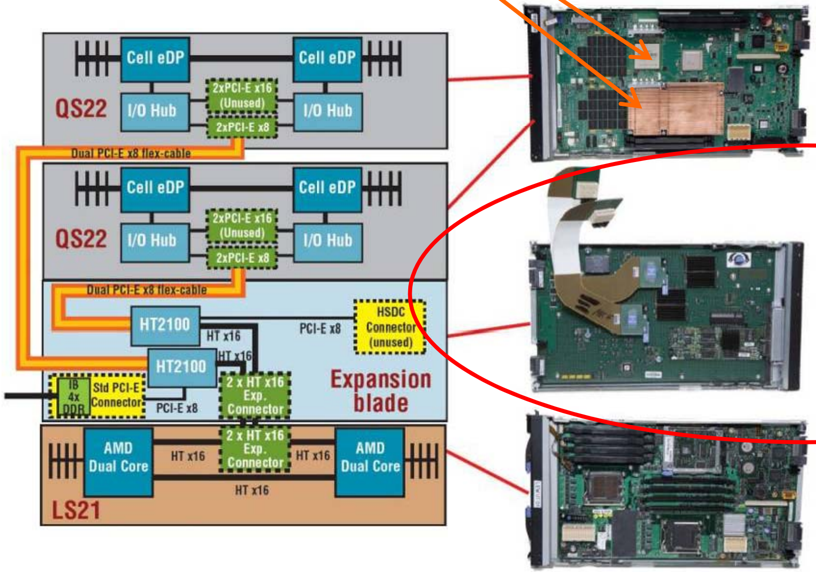
“Scalable Unit” Cluster Interconnect Switch/Fabric

Node-attached Cells is what makes Roadrunner different!

A Roadrunner TriBlade node integrates Cell and Opteron blades

New PowerXCell 8i chips
& QS22 blade





Two QS22's
with 2 Cells each

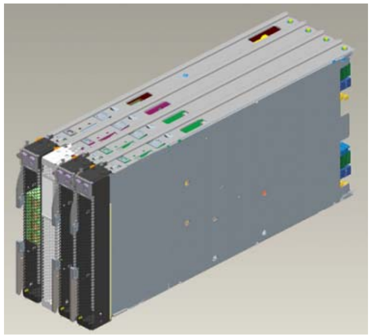
New part

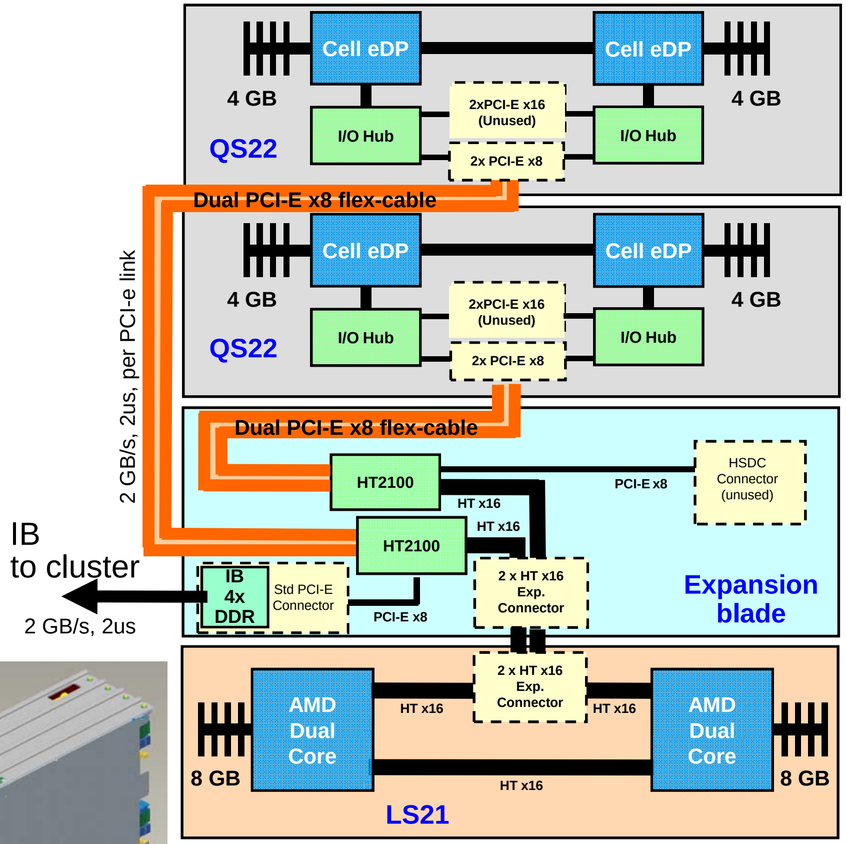
Expansion blade

LS21 with two
dual-core Opterons

A Roadrunner TriBlade node integrates Cell and Opteron blades

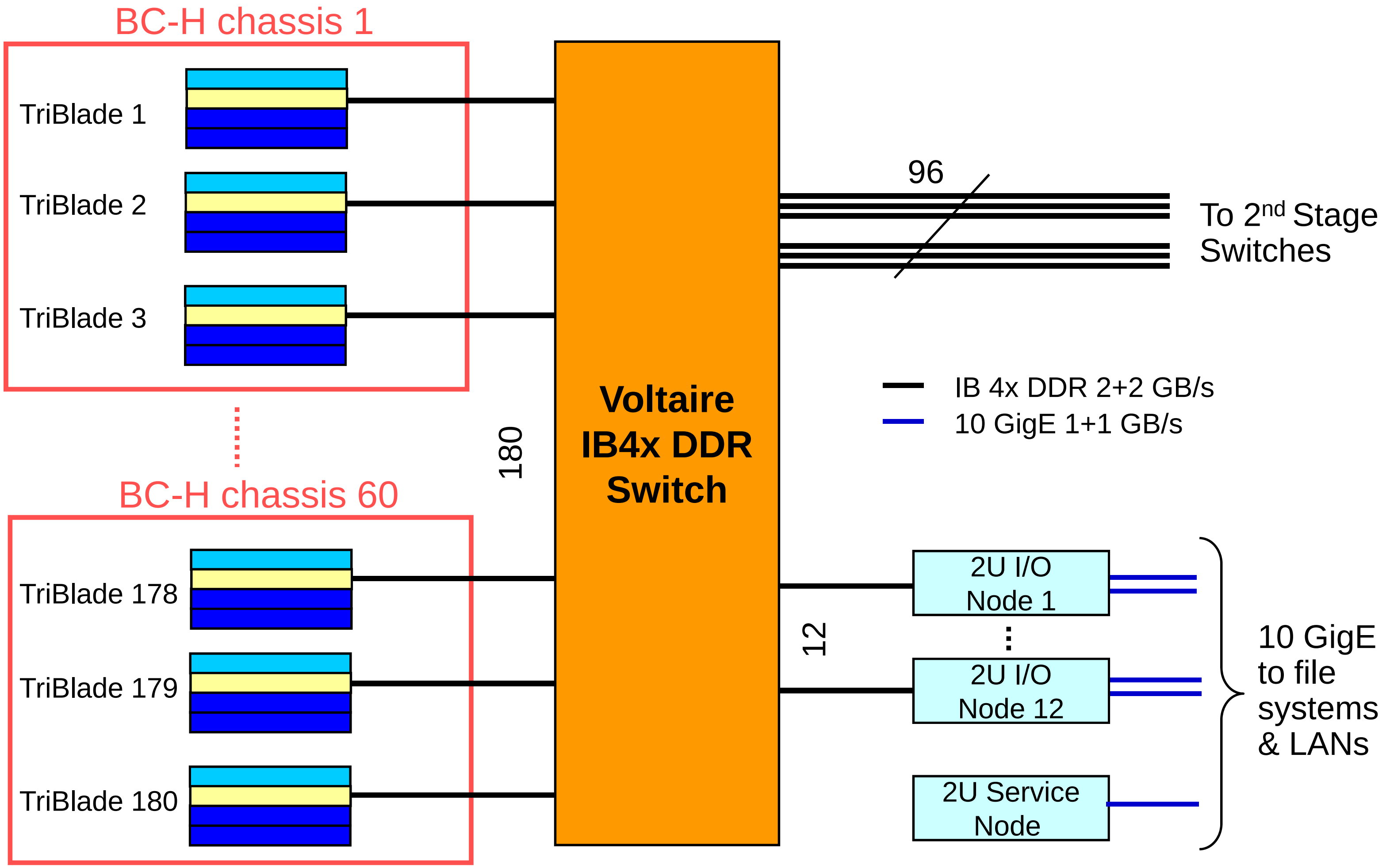
- QS22 is an IBM Cell blade containing two new enhanced double-precision (eDP/PowerXCell™) Cell chips
- Expansion blade connects two QS22 via four PCI-e x8 links to LS21 & provides the node's Mellanox ConnectX IB 4X DDR cluster attachment
- LS21 is an IBM dual-socket Opteron blade
- 4-wide IBM BladeCenter packaging
- Roadrunner Triblades are completely diskless and run from RAM disks with NFS & Panasas only to the LS21
- Node design points:
  - *One Cell chip per Opteron core*
  - *~400 GF/s double-precision & ~800 GF/s single-precision*
  - *16 GB Opteron memory PLUS 16 GB Cell memory*
  - *1 PCI-E x8 to each Cell*





Design point:
One Cell per Opteron core

A Connected Unit (CU) forms a building block



A Connected Unit (CU) is a powerful cluster

Connected Unit Specifications:

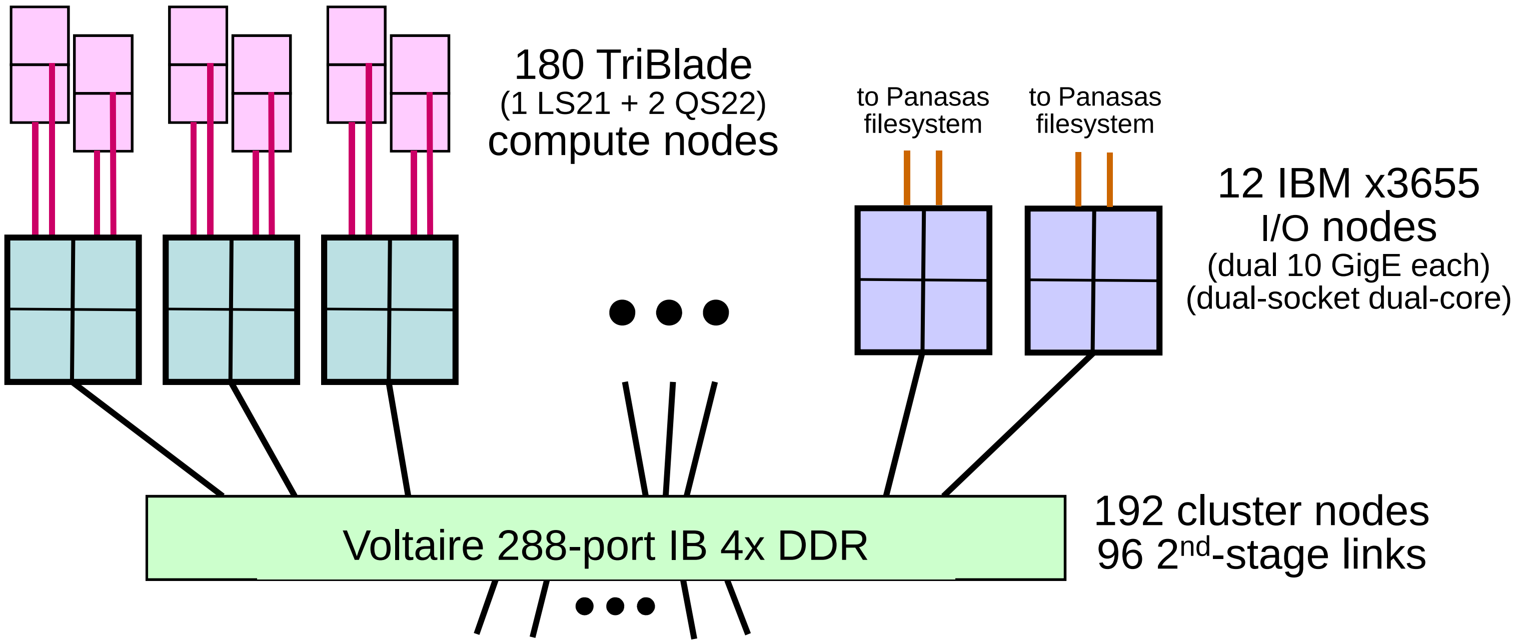
360 1.8 GHz dual-core Opterons
- 2.59 TF DP peak Opteron
- 2.88 TB Opteron memory

24 2.6 GHz dual-core Opterons
- in I/O nodes

720 PowerXCell chips
- **73.7 TF DP peak Cell**
- 2.88 TB Cell memory
- 18.4 TB/s Cell memory BW

192 IB 4X DDR cluster links
- 768 GB/s aggregate BW (bi-dir)
- 384 GB/s bi-section BW (bi-dir)

24 10 GigE I/O links on 12 I/O nodes
- 24 GB/s aggregate I/O BW (uni-dir)
  (IB limited)



180 TriBlade
(1 LS21 + 2 QS22)
compute nodes

to Panasas filesystem

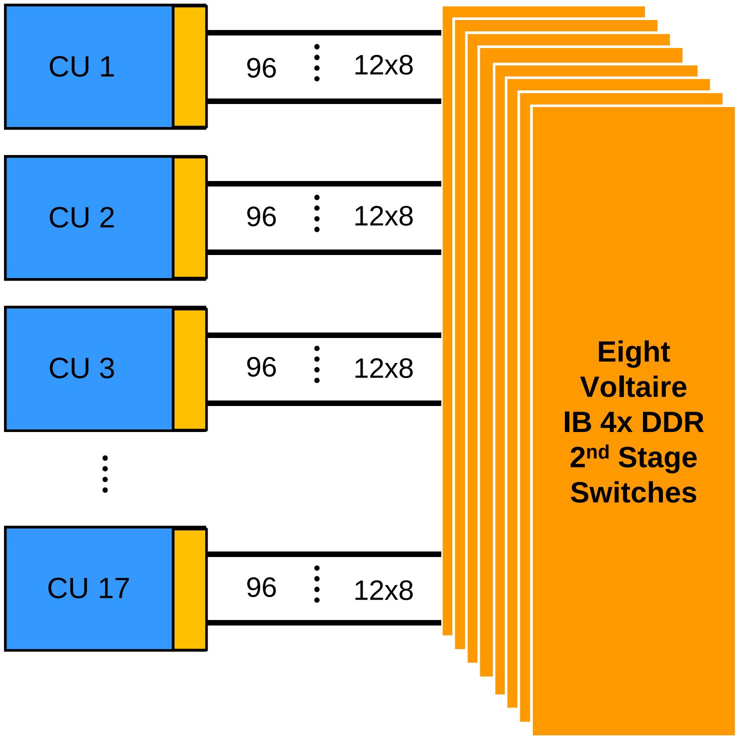
to Panasas filesystem

12 IBM x3655
I/O nodes
(dual 10 GigE each)
(dual-socket dual-core)

Voltaire 288-port IB 4x DDR

192 cluster nodes
96 2nd-stage links

Now build a cluster-of-clusters…



CU 1 96 ⋮ 12x8

CU 2 96 ⋮ 12x8

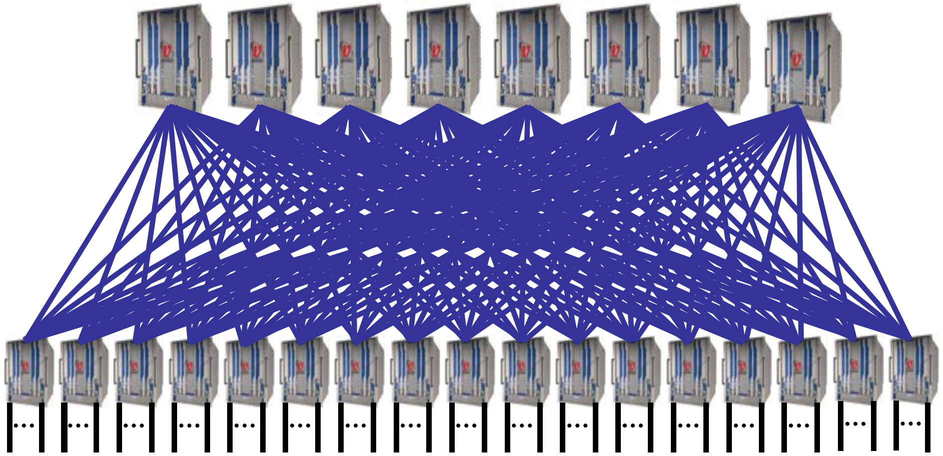
CU 3 96 ⋮ 12x8

⋮

CU 17 96 ⋮ 12x8

Eight Voltaire IB 4x DDR 2nd Stage Switches

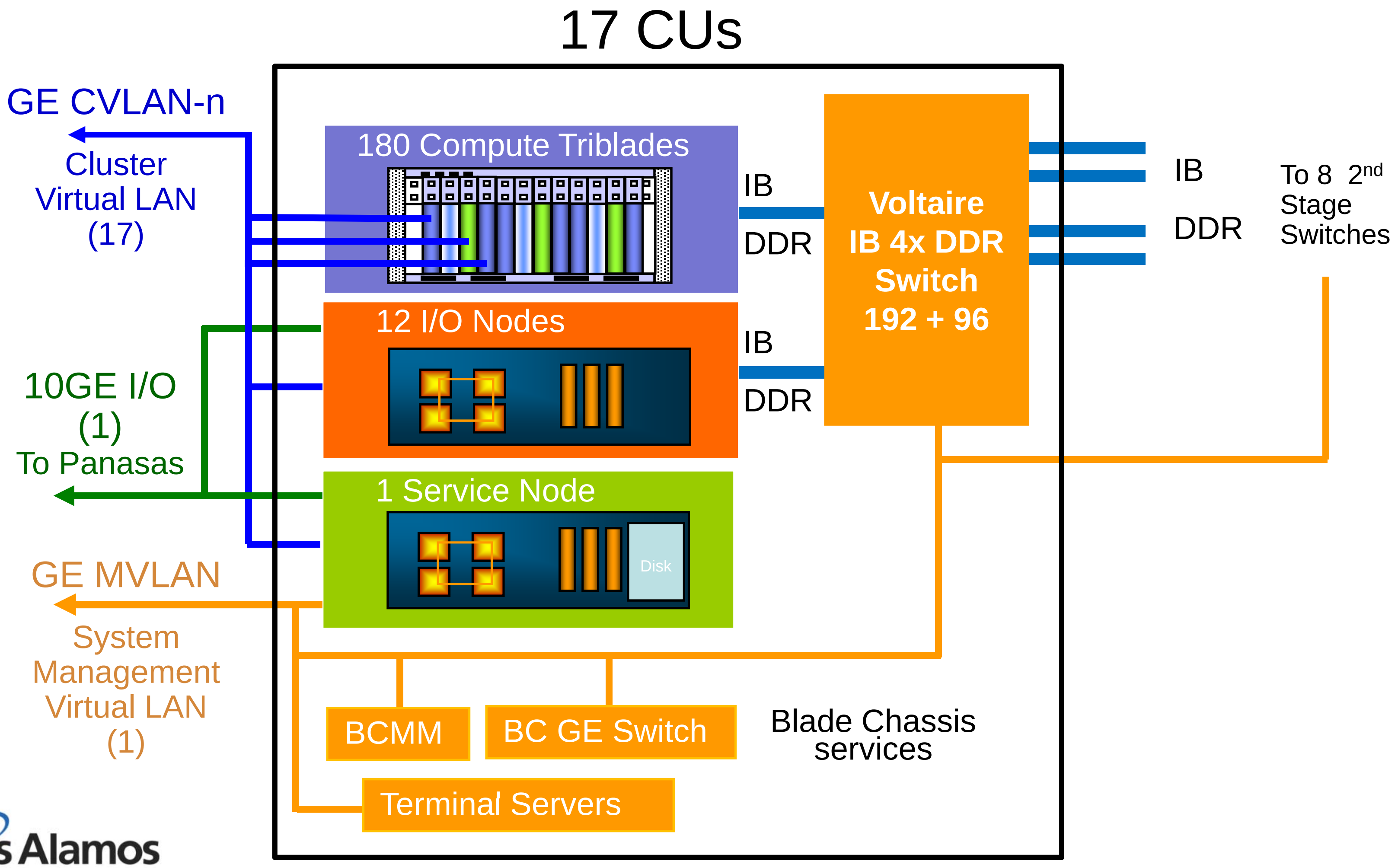
2nd–stage switches form a half-bandwidth fat-tree



17 CUs with CU switches, 3264 IB nodes

Extra 2nd–stage switch ports allow expansion up to 24 CUs

Roadrunner System Networks



Roadrunner is a petascale system in 2008

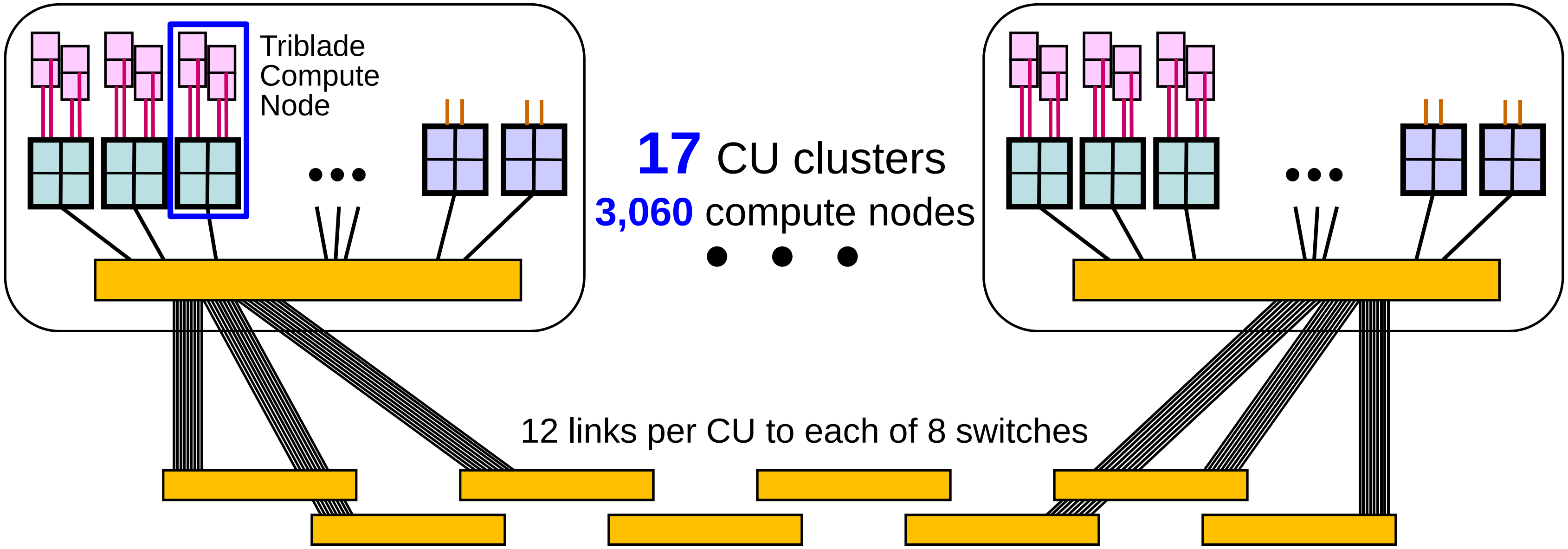
3,060 Compute Nodes
- 6,120 dual-core Opterons
- 44.1 TF DP peak Opteron
- 49 TB Opteron memory

204 I/O nodes w/ 408 Opterons

12,240 PowerXCell 8i chips
- 1.33 PF DP peak Cell
- 2.59 PF SP peak Cell
- 49 TB Cell memory
- 313 TB/s Cell memory BW

2-stage IB 4X DDR interconnect
- 13.1 TB/s aggregate BW (bi-dir) (1st stage)
- 6.5 TB/s aggregate BW (bi-dir) (2nd stage)
- 3.3 TB/s bi-section BW (bi-dir) (2nd stage)

204 I/O nodes with 408 10 GigE links to a Panasas parallel file system



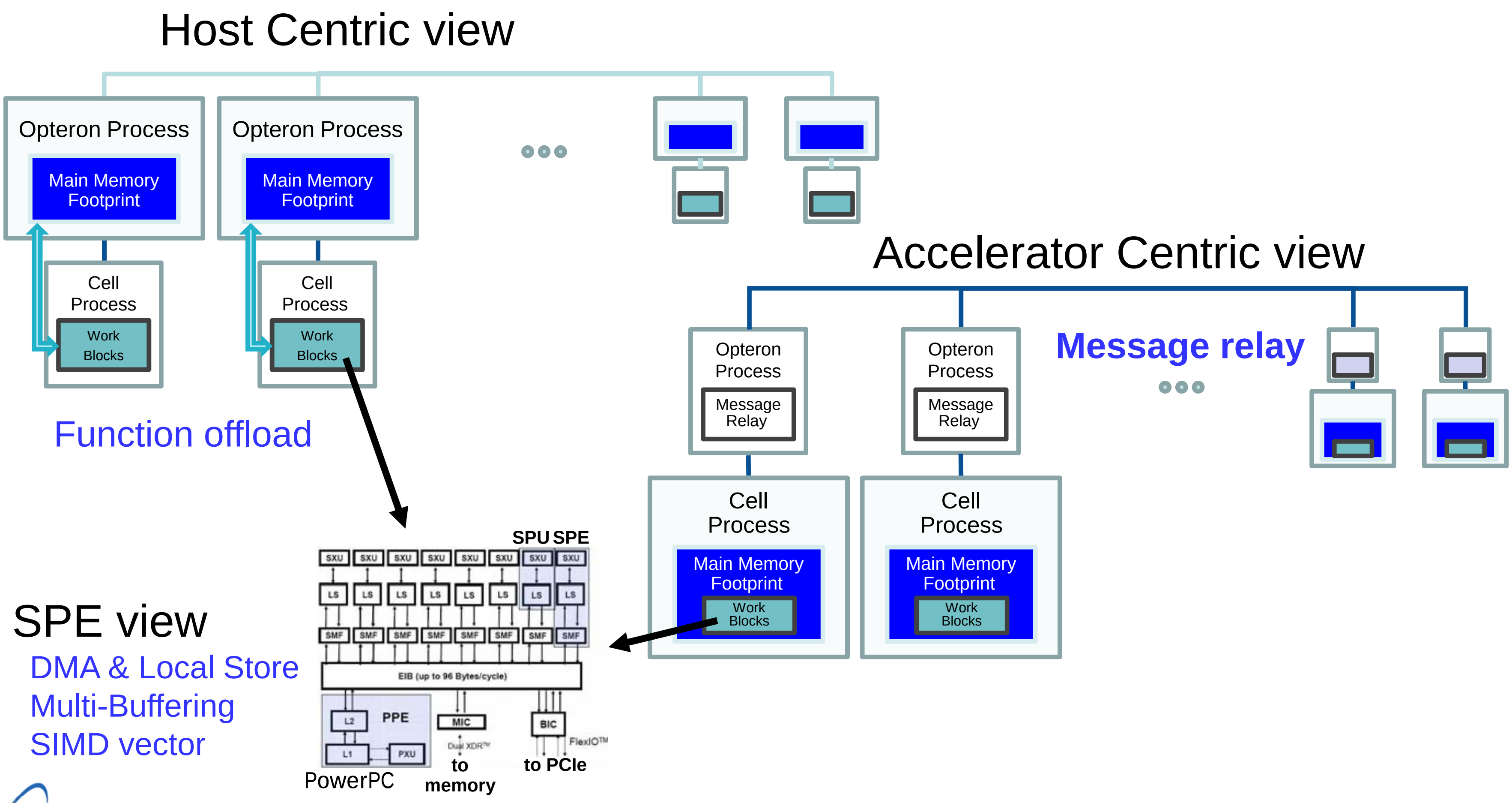
Eight 2nd-stage IB 4X DDR switches

Roadrunner at a glance

- **Cluster of 17 Connected Units (CU)**
  - *12,240 IBM PowerXCell 8i chips*
  - *1.33 Petaflop/s DP peak (Cell)*
  - *1.026 PF sustained Linpack (DP)*
  - *6,120 (+408) AMD dual-core Opterons*
  - *44.1 (+4.4) Teraflop/s peak (Opteron)*
- **InfiniBand 4x DDR fabric**
  - *3264 total nodes; 2-stage fat-tree; all-optical cables*
  - *Full bi-section BW within each CU*
    - 384 GB/s (bi-directional)
  - *Half bi-section BW among CUs*
    - 3.26 TB/s (bi-directional)
- **~100 TB aggregate memory**
  - *49 TB Opteron (compute nodes)*
  - *49 TB Cell*
- **~200 GB/s sustained File System I/O:**
  - *204x2 10GE Ethernets to Panasas*
- **Fedora Linux**
  - *On LS21 & QS22 blades & I/O & service nodes*
- **SDK for Multicore Acceleration**
  - *Cell compilers, libraries, tools*
- **xCAT Cluster Management**
  - *System-wide GigEnet network*
- **2.35 MW Power:**
  - *0.437 GF/Watt*
- **Area:**
  - *280 racks*
  - *5200 ft^2*

Programming Concepts

Programming Approaches for Roadrunner



Host Centric view

Opteron Process

Main Memory Footprint

Cell Process

Work Blocks

Function offload

Accelerator Centric view

Opteron Process

Message Relay

Cell Process

Main Memory Footprint

Work Blocks

Message relay

SPE view

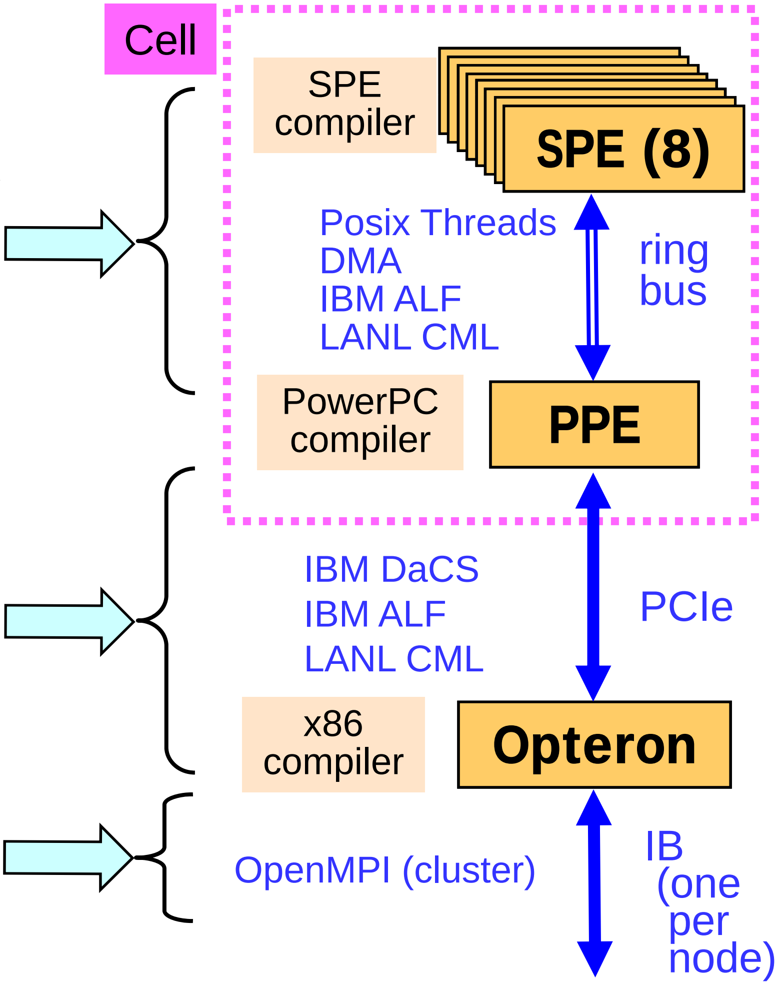
DMA & Local Store

Multi-Buffering

SIMD vector

Three types of processors work together

- Parallel computing on Cell
  - *data partitioning & work queue pipelining*
  - *process management & synchronization*
- Remote communication to/from Cell
  - *data communication & synchronization*
  - *process management & synchronization*
  - *computationally-intense offload*
- MPI remains as the foundation



Cell

SPE compiler

SPE (8)

Posix Threads
DMA
IBM ALF
LANL CML

ring bus

PowerPC compiler

PPE

IBM DaCS
IBM ALF
LANL CML

PCIe

x86 compiler

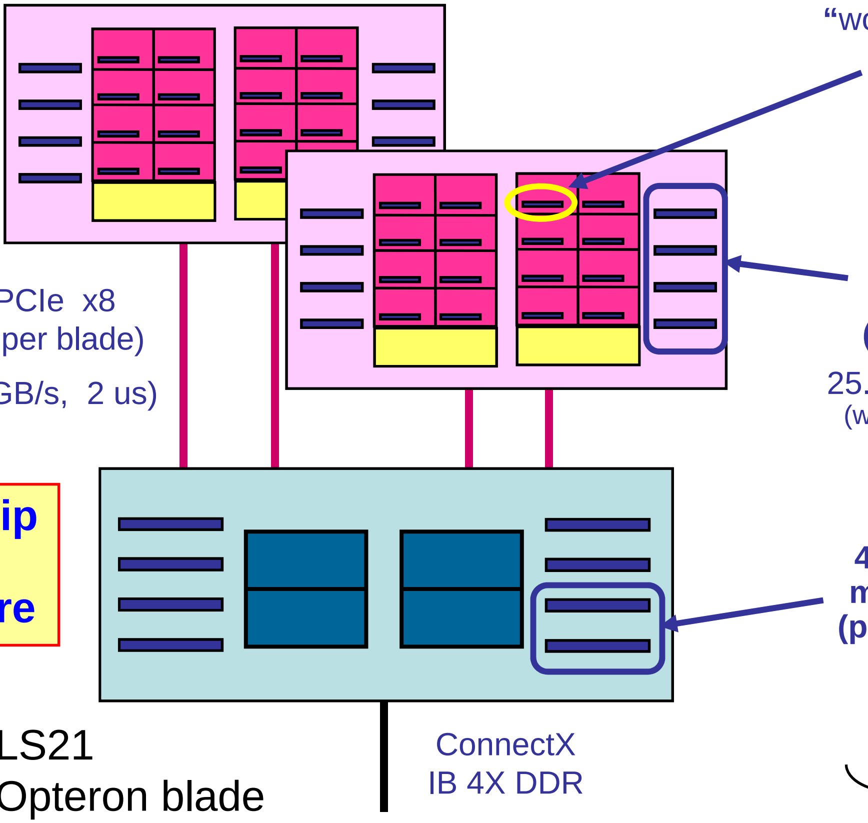
Opteron

OpenMPI (cluster)

IB (one per node)

Roadrunner nodes have a memory hierarchy

QS22 Cell blades



PCIe x8
(2 per blade)

(2 GB/s, 2 us)

One Cell chip per Opteron core

LS21
Opteron blade

ConnectX
IB 4X DDR

(2 GB/s, 2 us)

256 KB of
"working" memory
(per SPE)

25.6 GB/s
off-SPE BW

~200 GB/s per
Cell on EIB bus

4 GB of shared memory (per Cell)

25.6 GB/s/chip
(w/ 800 DDR2)

8 GB of NUMA shared memory (per blade)

16 GB of distributed memory (per node)

4 GB of memory (per core)

8 GB of shared memory (per socket)

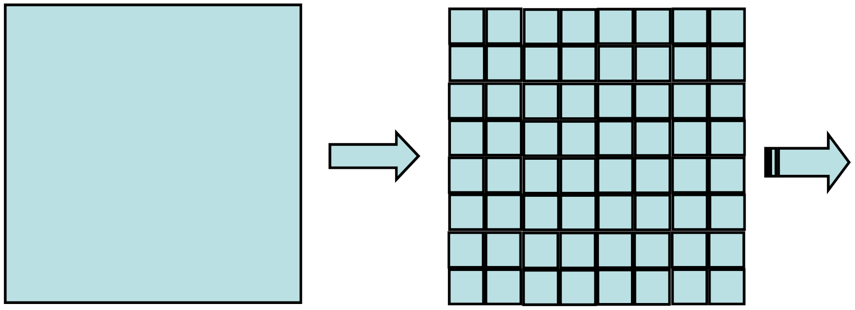
5.4 GB/s/core

16 GB of NUMA shared memory (per node)

"equal memory size" concept

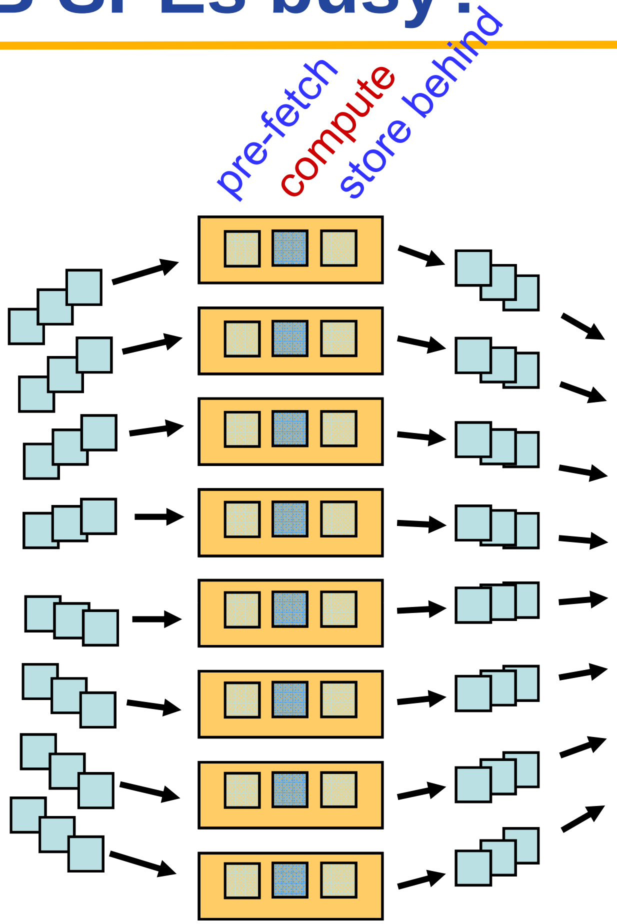
How do you keep the 256KB SPEs busy?

Break the work into a stream of pieces



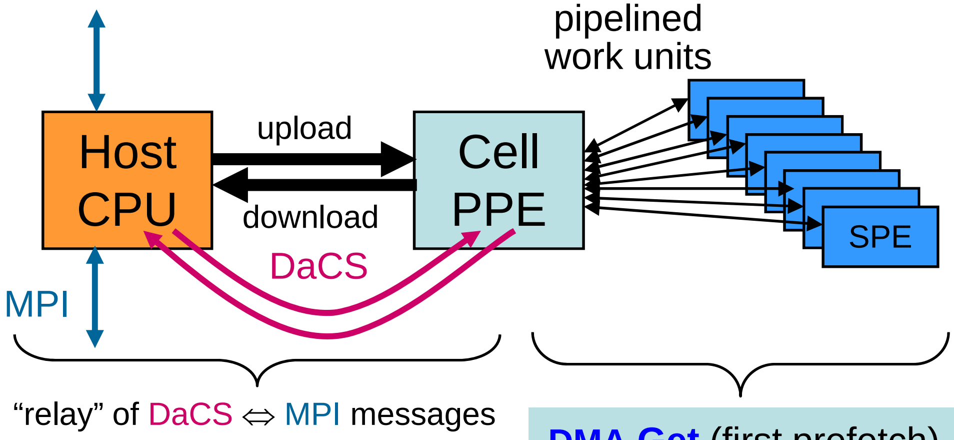
problem domain of a Cell processor

grid tiles or particle bundles (can include ghost zones)



data chunks stream in & out of 8 SPEs using asynch DMAs and triple-buffering

Put it all together: MPI+DaCS+DMA+SIMD



Compute & memory DMA transfers are overlapped in HW!

MPI & DaCS can also be fully asynchronous

DMA Get (first prefetch)
Switch work buffers

DMA Get (prefetch)
DMA Wait (complet current)
Compute
DMA Put (store behind)
DMA Wait (previous put)
Switch work buffers

DMA Wait (put)

- DMAs are simply block memory transfers
  - *HW asynchronous (no SPE stalls)*
  - *DDR2 memory latency and BW performance*

DMA Get:

```
  mfc_get( LS_addr, Mem_addr, size, tag, 0, 0);
```

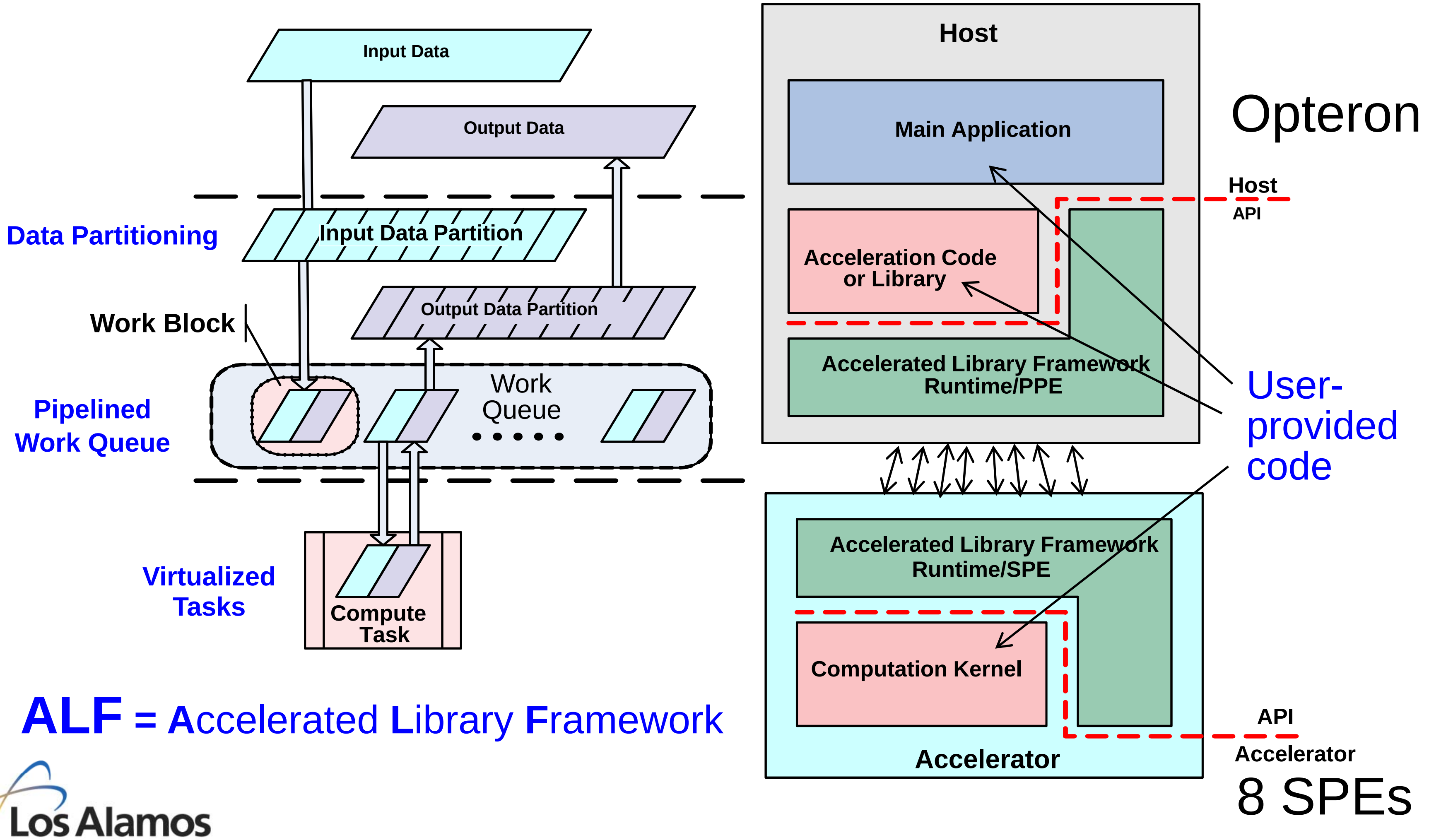
DMA Put:

```
  mfc_put( Mem_addr, LS_addr, size, tag, 0, 0);
```

DMA Wait:

```
  mfc_write_tag_mask(1<<tag);
  mfc_read_tag_status_all();
```

IBM-ALF is a simple work-queue approach for abstracting parallelism directly to SPEs



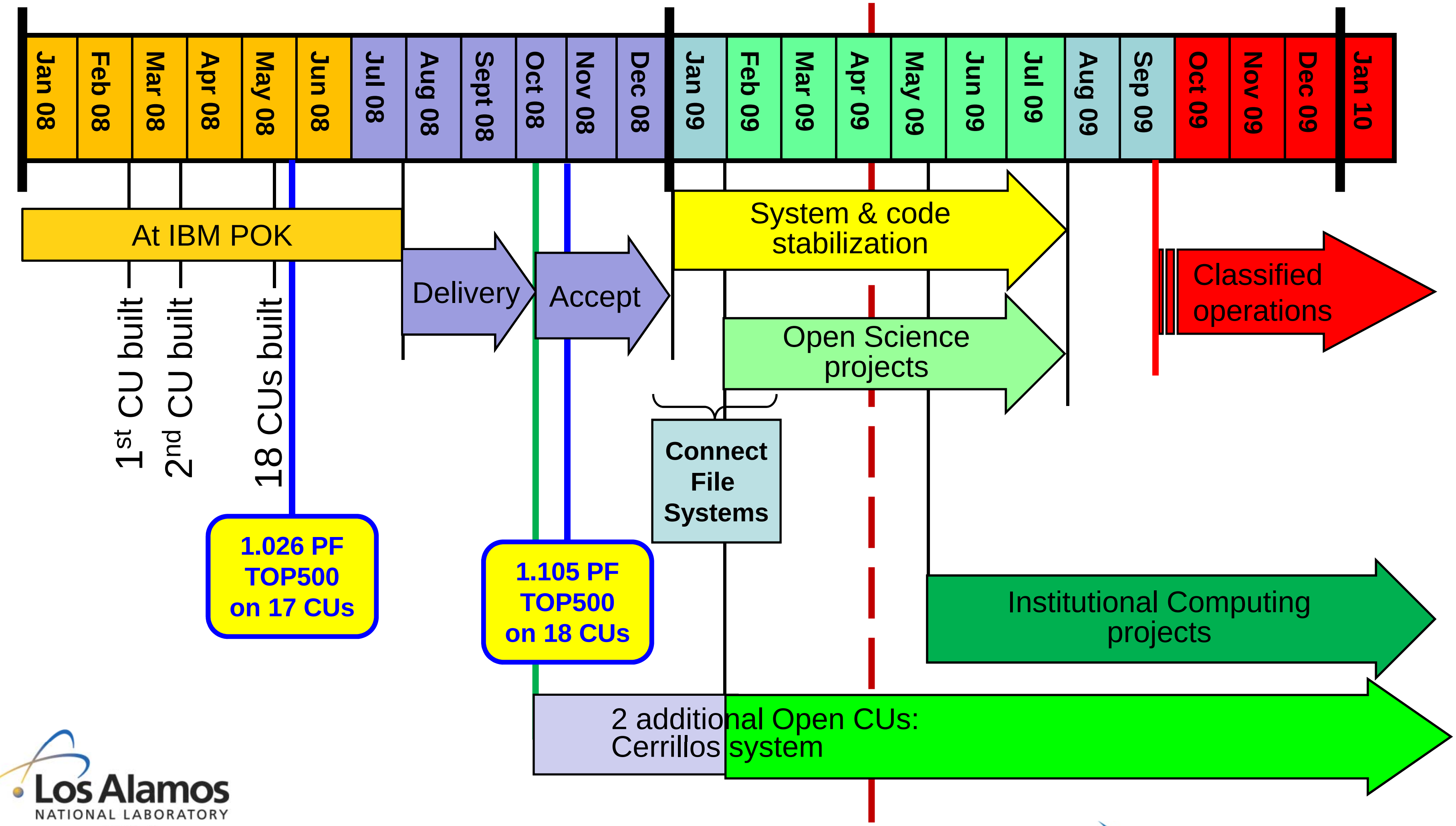
ALF = **A**ccelerated **L**ibrary **F**ramework

Programming approach has now been demonstrated and is Tractable

- Two levels of parallelism:
  - *node-to-node: MPI & DaCS-MPI-DaCS relay*
  - *within-Cell: threads, pipelined DMAs, & SIMD*
- Large-grain computationally intense portions of code are split off for Cell acceleration within a node process
  - *Usually an entire tree of subroutines*
  - *This is equivalent to "function offload" of entire large algorithms*
- Threaded fine-grained parallelism introduced within the Cell itself
  - *Create many-way parallel pipelined work units for the 8 SPEs*
  - *Good for both multicore/manycore chips and heterogeneous chip trends with dwindling memory bandwidth*
- Communications during Cell computation are possible between Cells via DaCS-MPI-DaCS relay approach
- Considerable flexibility and opportunities exist beyond this approach

Five Waves of Roadrunner Applications Codes

Recent Roadrunner History



1.026 PF TOP500 on 17 CUs

1.105 PF TOP500 on 18 CUs

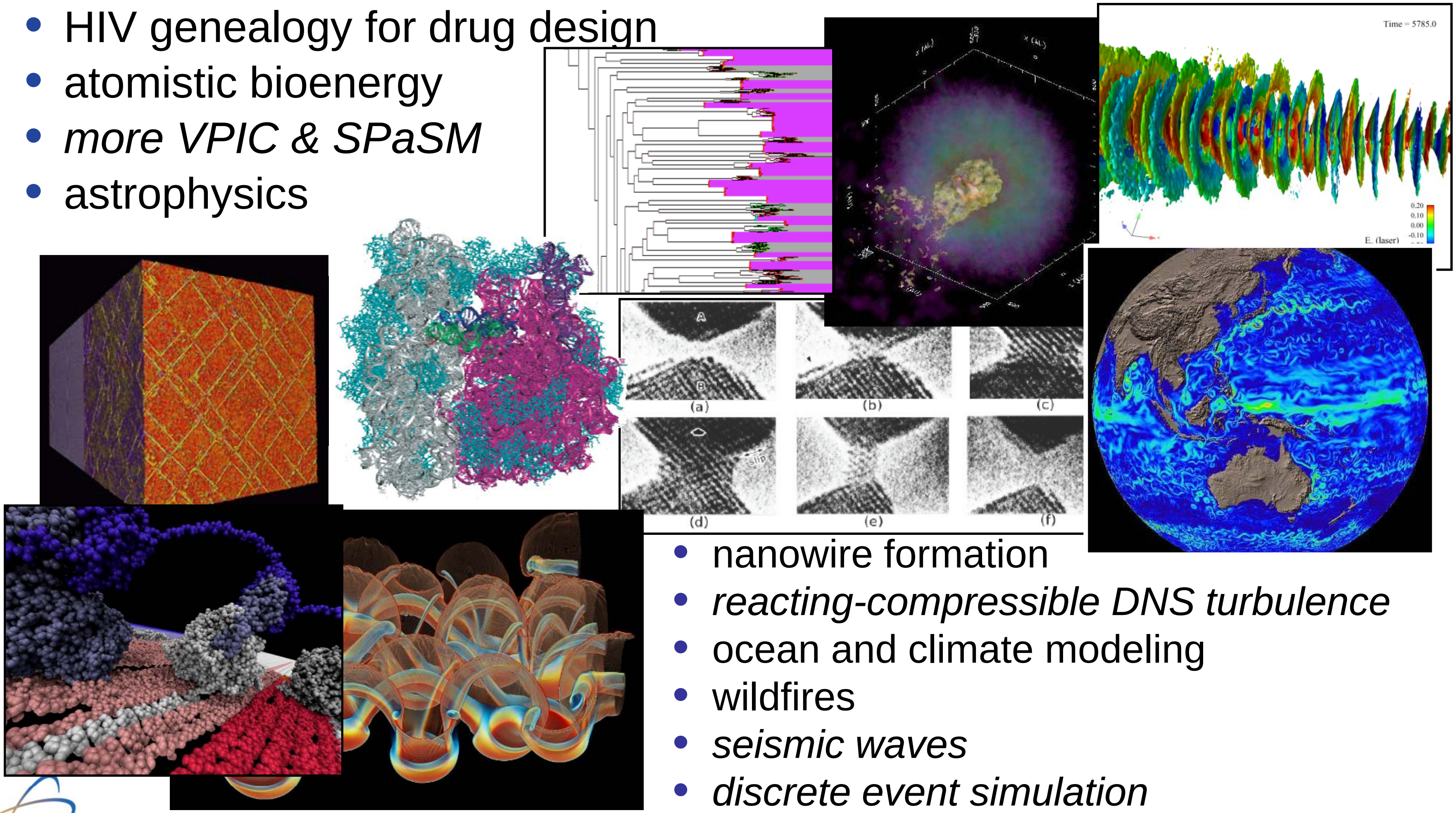
Five Waves of Roadrunner Application Codes

1. Assessment Codes (Oct. 2006 – Oct. 2007)
   - *Proof of Cell & Hybrid programming capability: 4 codes*
   - *Prototype hardware: old Cell/QS20 blades & very first eDP Cell*
2. Full-System Pre-Acceptance Testing (June – Nov. 2008)
   - *Gordon Bell finalists: VPIC & SPaSM*
   - *PetaVision (sustained 1+ single-precision-PF!)*
   - *PPM (Paul Woodward)*
3. Roadrunner Open Science (Oct. 2008 – July 2009)
   - *8 codes & 10 projects*

   today
4. Institutional Computing (starting May 2009 on Cerrillos – 2 CUs)
   - *19 new projects to start*
5. Classified ASC Use (starting Oct. 2009)

Roadrunner Open Science and Institutional Computing

- HIV genealogy for drug design
- atomistic bioenergy
- *more VPIC & SPaSM*
- astrophysics



- nanowire formation
- *reacting-compressible DNS turbulence*
- ocean and climate modeling
- wildfires
- *seismic waves*
- *discrete event simulation*
- *piecewise-parabolic method (PPM)*

Exciting opportunities among the 10 selected proposals for Roadrunner Open Science

Kinetic Thermonuclear Burn Studies with VPIC on Roadrunner	VPIC
Multibillion-Atom Molecular Dynamics Simulations of Ejecta Production and Transport using Roadrunner	SPaSM
New frontiers in viral phylogenetics	ML
Three-Dimensional Dynamics of Magnetic Reconnection in Space and Laboratory Plasmas	VPIC
The Roadrunner Universe	MC^3
Implicit Monte Carlo Calculations of Supernova Light-Curves	IMC + Rage
Instabilities-Driven Reacting Compressible Turbulence	CFDNS
Cellulosomes in Action: Peta-Scale Atomistic Bioenergy Simulations	GROMACS
Parallel-replica dynamics study of tip-surface and tip-tip interactions in atomic force microscopy and the formation and mechanical properties of metallic nanowires	PAR-REP + CellMD
Saturation of Backward Stimulated Scattering of Laser In The Collisional Regime	VPIC

🟩 Indicates new work ▩ Indicates new + old

The LANL Roadrunner web site is

http://www.lanl.gov/roadrunner/

Roadrunner architecture
Early applications efforts
Upcoming Open Science efforts
Cell & hybrid programming
Computing trends
Related Internet links