Roadrunner Platform Overview

Roadrunner Technical Seminar Series
March 13, 2008

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!

LA-UR-08-1994

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 Cells*
- Concepts for Programming Roadrunner
  - *Something old, something new*
  - *MPI, Opteron+Cell, local memory & DMAs*

Ways to build a Petaflop/s supercomputer today

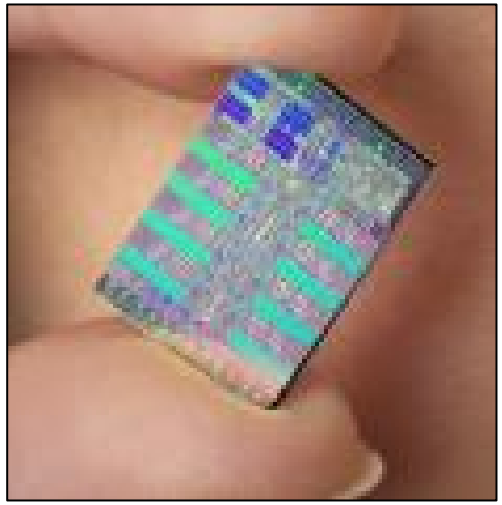
- Opteron cluster (e.g. ~2X Ranger/TACC)
  - *32,000 quad-core Opterons (8,000 nodes, 130K cores)*
- Cray XT3/4 (e.g. Baker/ORNL sooner)
  - *32,000 quad-core Opterons (16,000 MPP nodes, 130K cores)*
- IBM BlueGene/P (bigger sooner)
  - *80,000 BG/P PPC processors (80,000 MPP nodes, 320K cores)*
- IBM Cell-accelerated Roadrunner cluster
  - *10,000 Cells (2,500 nodes, 80K Cell SPUs)*
- There are advantages and disadvantages of each approach!

Roadrunner is specifically about a unique future trend

Some background

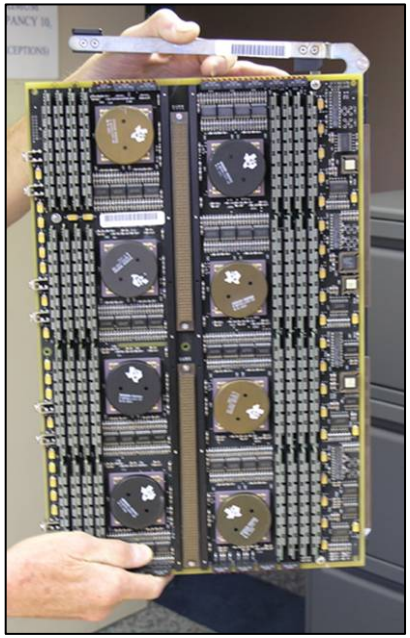
Roadrunner project is a partnership with IBM

- Contract contract signed September 8, 2006 with IBM
- Roadrunner has 3 phases
  - ***Phase 1*** *(Base system) supports near-term mission deliverables*
  - ***Phase 2*** *(Hybrid-node prototype) supports pre-Final assessment*
  - ***Phase 3*** *(Hybrid final system)*
    - Achieves PetaFlops level of performance
    - Demonstrates new paradigm for high performance computing
- Accelerated vision of the future
  - *New programming paradigm*
  - *Faster processing on accelerators*
  - *Still leveraging the marketplace*



Cell processor (2007, 100 GF)

100x in
14 yrs
8 vector units each



CM-5 board (1994, 1 GF)

Our vision for high-performance computing embraces both multi- and unit-physics codes

Advanced Architecture for algorithms and applications

- *Target **select** physics implementations and convert*
- *Provide faster solutions or improved accuracy in key areas*
- *Incrementally update existing integrated codes for targeting key uncertainties*
  - Target focused simulations, not general usage
  - Focus is more predictive science oriented than speeding up production jobs

Roadrunner will lay the groundwork for advanced architectures that can significantly increase simulation performance in the future

Multi-scale, multi-physics codes
Complex & varied physics models
100x slower than Unit-physics codes
Under-resolved

Roadrunner will provide resources necessary for predictive science at scale simulations

Experiments (NIF, ZR …)
Complex & varied physics
Difficult @ high-energy-density
Under-resolved diagnostics
Tend to be integral

Unit-physics codes
Few physics assumptions
100x faster than Multi-physics
Fully resolved
Simple, repetitive manipulations

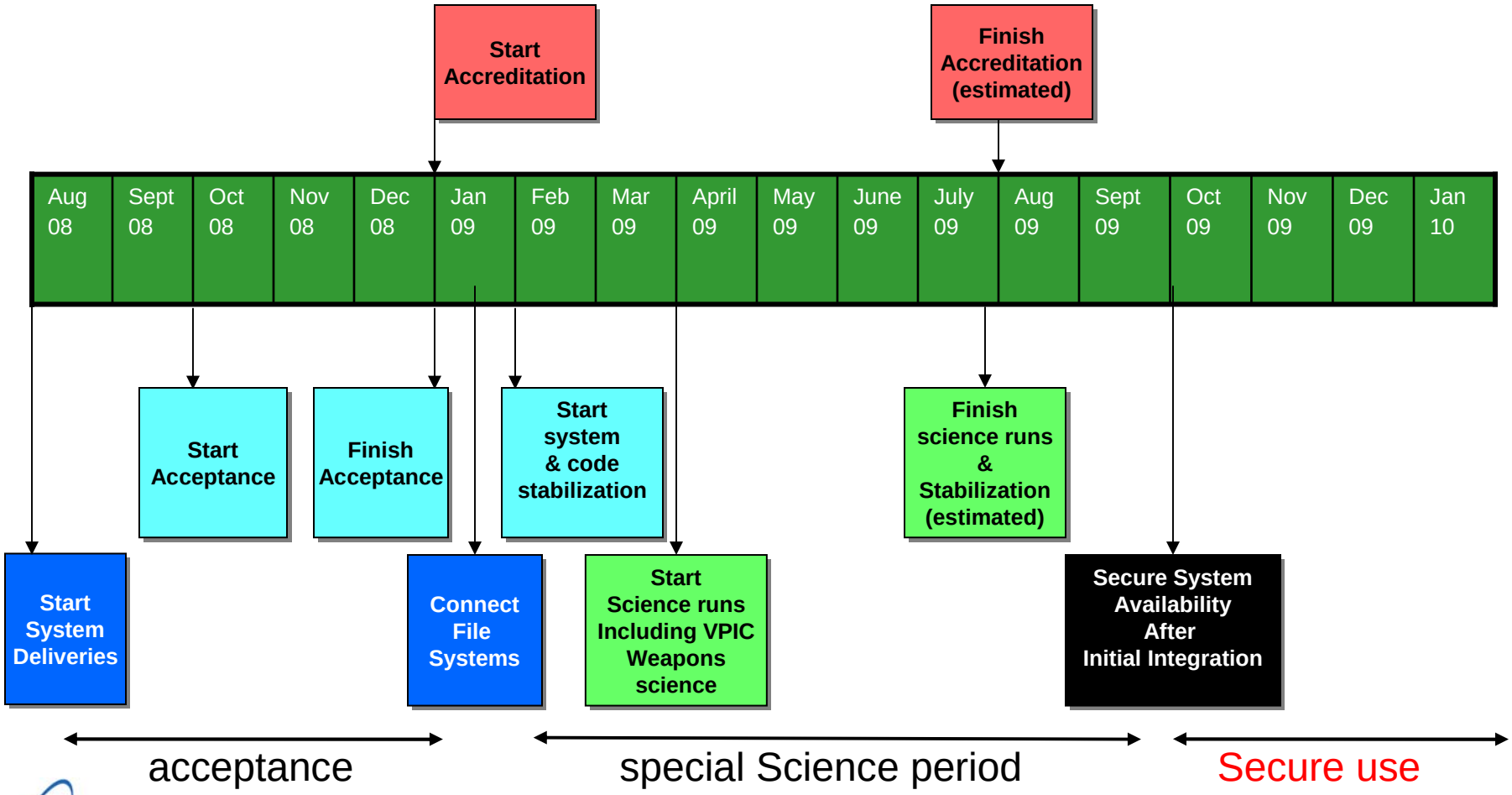
Science @ Scale

- *Multi-scale unit physics for weapons & open **science***
  - Validate model assumptions & better understand physics
  - Cross-validate physics models at overlapping resolutions
- *Run at Petascale (machine and phenomenological scales)*

Roadrunner Final System Technical Assessment charter

- (1) Assess the potential for LANL and IBM to achieve the following objectives for the final accelerated Roadrunner system during its useful lifetime, and
  (2) provide suggestions for improving the prospects for success.
  - *IA. Provide effective computing resources for important science at scale simulations beginning in FY09.*
  - *IB. Provide an advanced architecture that can significantly increase application performance in the future.*
  - *II. Achieve a sustained petaflop/s on LINPACK in FY08.*
  - *III. Develop a programming model and provide a reasonable Application Programming Interface (API) for programming hybrid systems.*
  - *IV. Effectively manage and integrate the final system at operational scale.*
  - *V. Develop appropriate new technology and deliver the Roadrunner final system to Los Alamos in Summer 2008.*
- **This Roadrunner Technical Seminar Series will present the efforts that addressed these questions**
  - ***What was technically accomplished and how?***

Roadrunner schedule at LANL



A Roadrunner is born

Roadrunner is being built up now at IBM

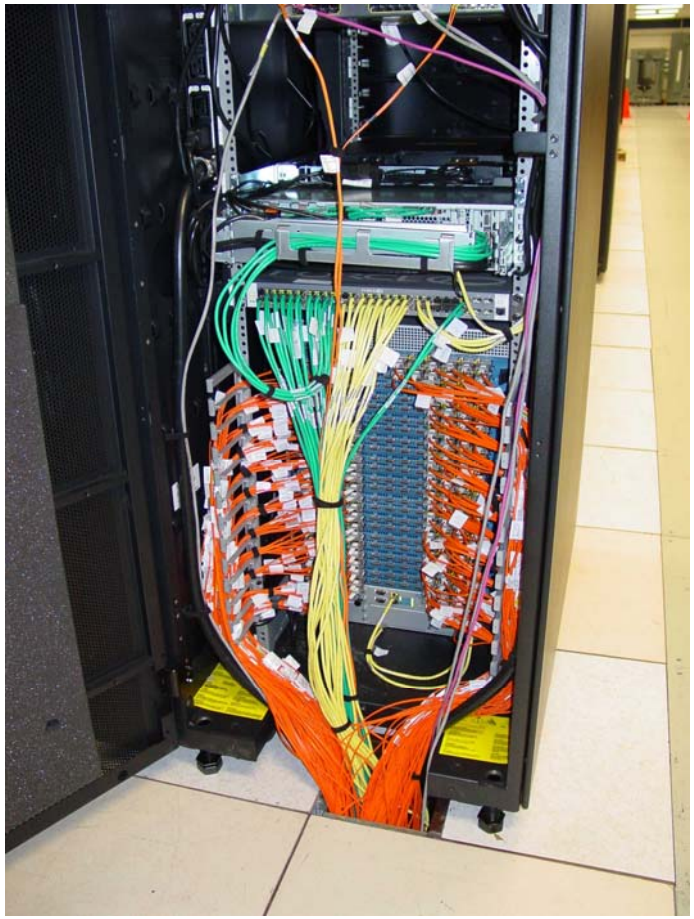
The very first Roadrunner racks operating at IBM-POK

Preparing the next part of Roadrunner

Preparing racks and stringing wires

Roadrunner uses thousands of parts

InfiniBand optical cables
and 1-of-26 switches

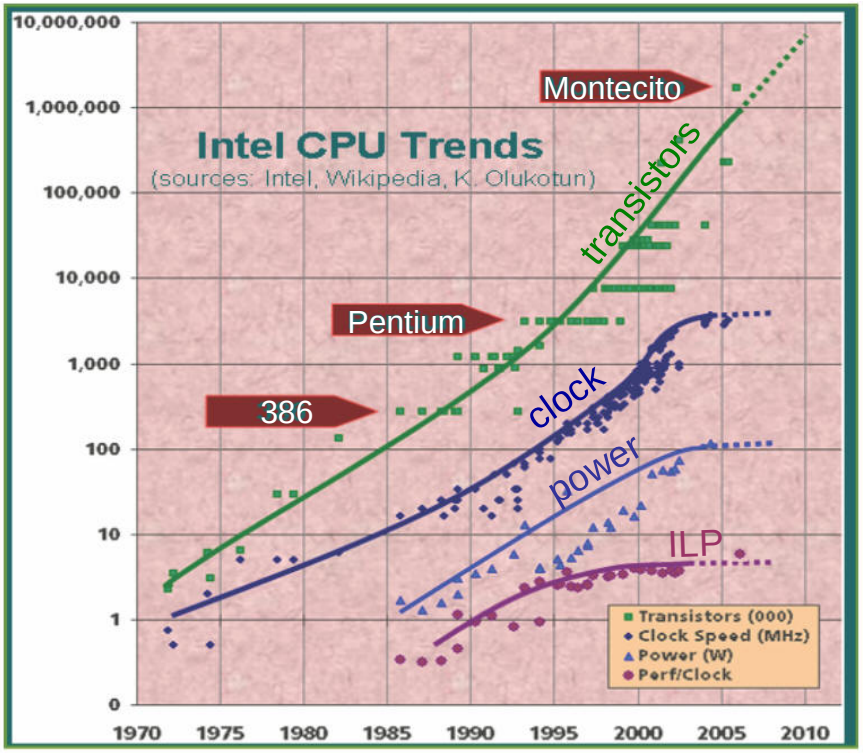


The Cell Processor

a harbinger of the future

Microprocessor trends are changing

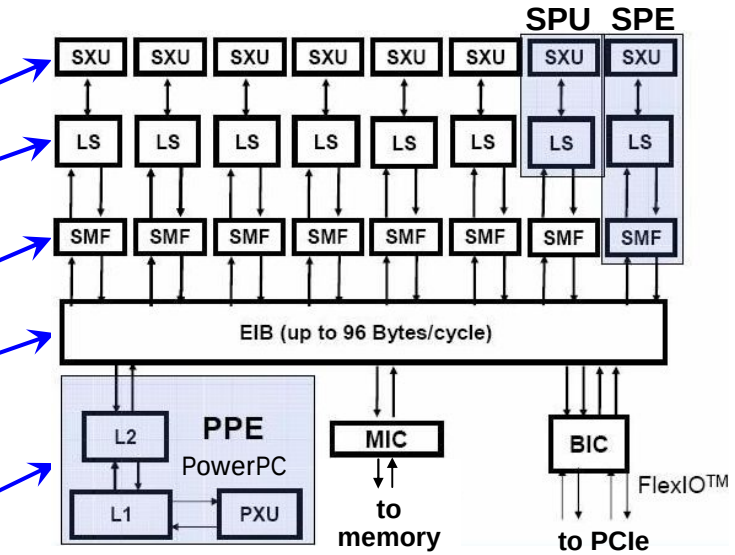
- Moore's law still holds, but is now being realized differently
  - *Frequency, power, & instruction-level-parallelism (ILP) have all plateaued*
  - *Multi-core is here today and many-core (≥ 32) looks to be the future*
  - *Memory bandwidth and capacity per core are headed downward (caused by increased core counts)*
  - *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*

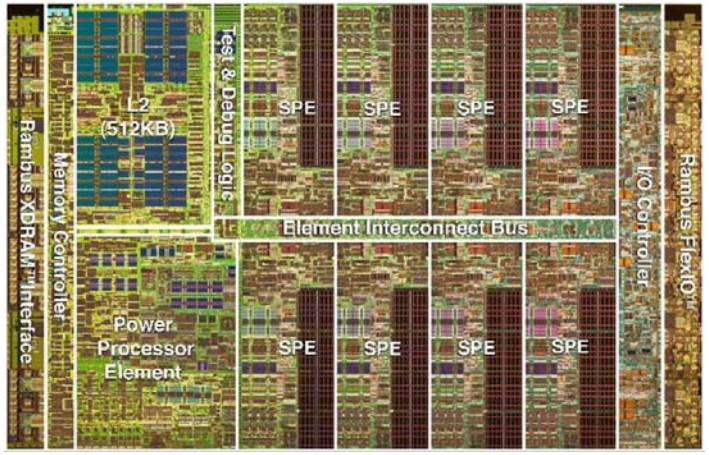


From Burton Smith, LASCI-06 keynote, with permission

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

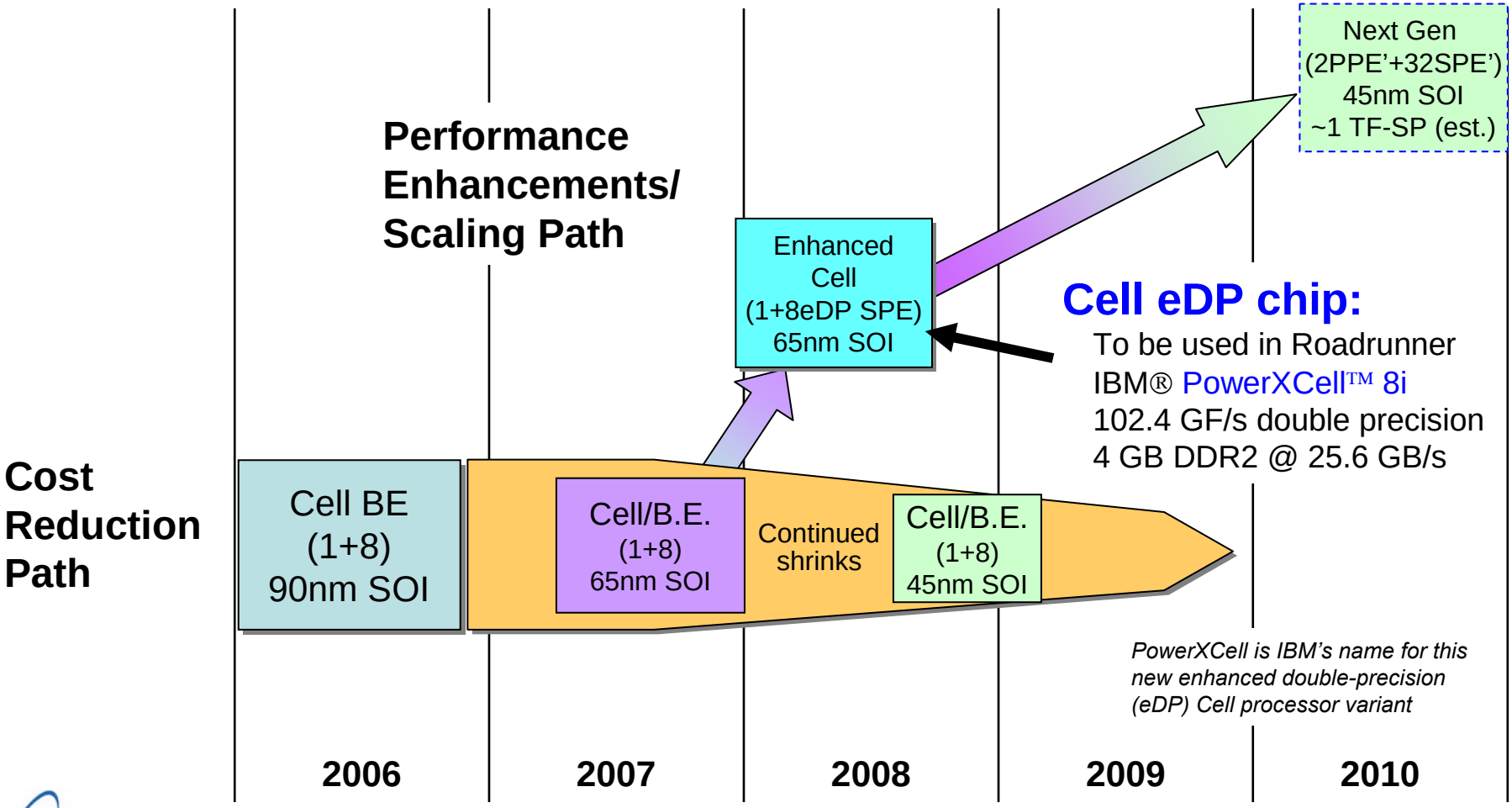
- Cell Broadband Engine (CBE*) developed by Sony-Toshiba-IBM
  - used in Sony PlayStation 3
- 8 Synergistic Processing Elements (SPEs)
  - 128-bit vector engines
  - 256 kB local memory (LS = Local Store)
  - Direct Memory Access (DMA) engine (25.6 GB/s each)
  - Chip interconnect (EIB)
  - Run SPE-code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





* trademark of Sony Computer Entertainment, Inc.

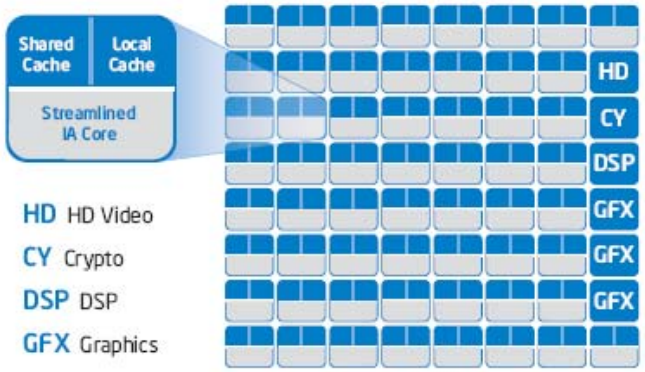
IBM is creating new Cell processors



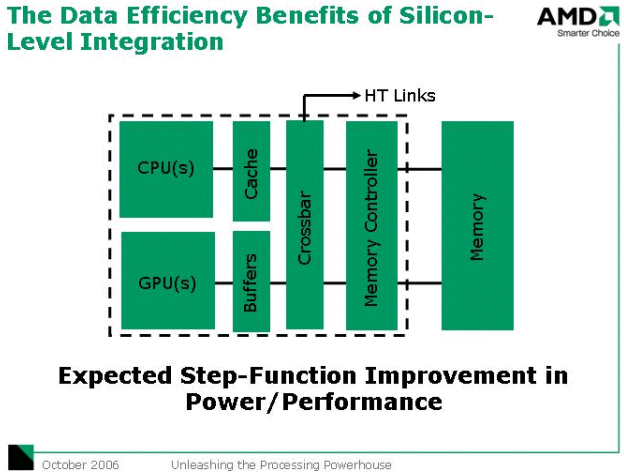
All future dates and specifications are estimations only; Subject to change without notice.
Dashed outlines indicate concept designs.

Industry presentations show changing trends in processors

Intel's Microprocessor Research Lab



AMD Fusion

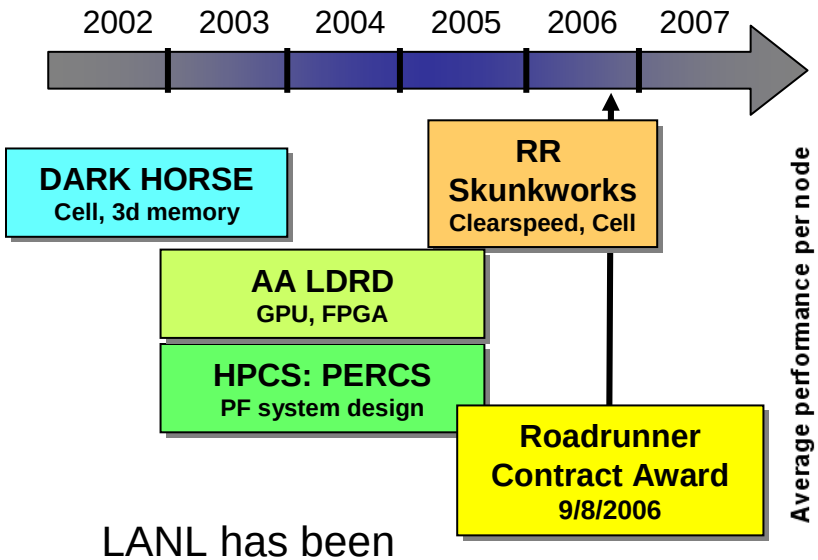


Intel's Visual Computing Group - Larabee

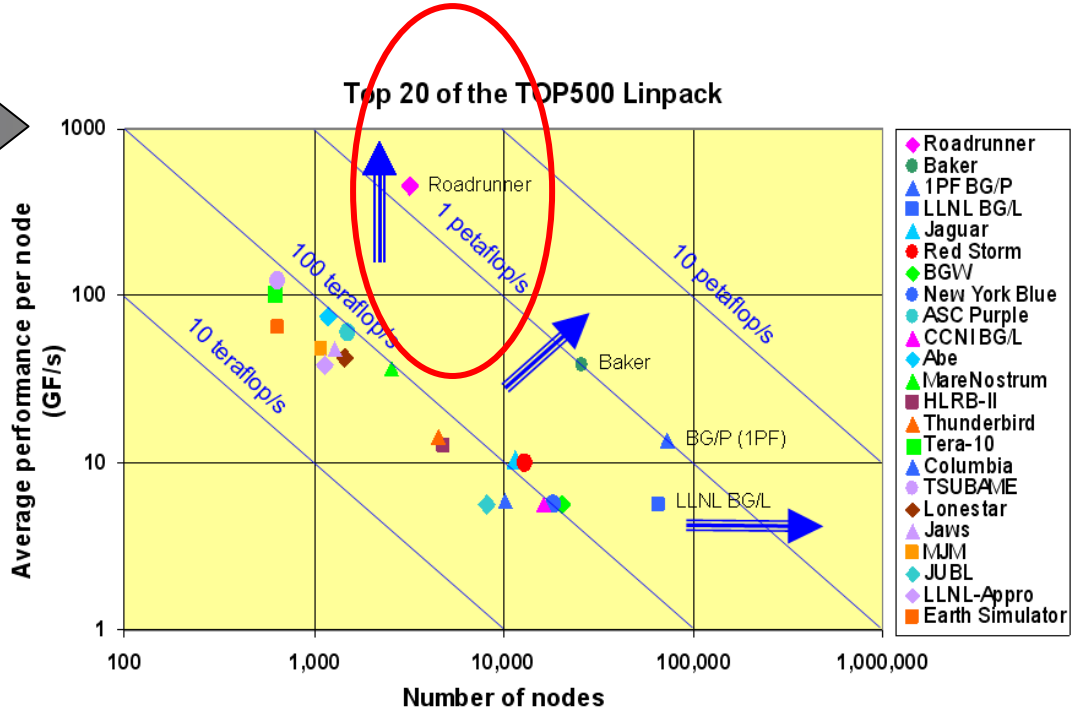
nVidia G80 - 2006

Taken from publicly available information

Hybrid computing is a transformational technology



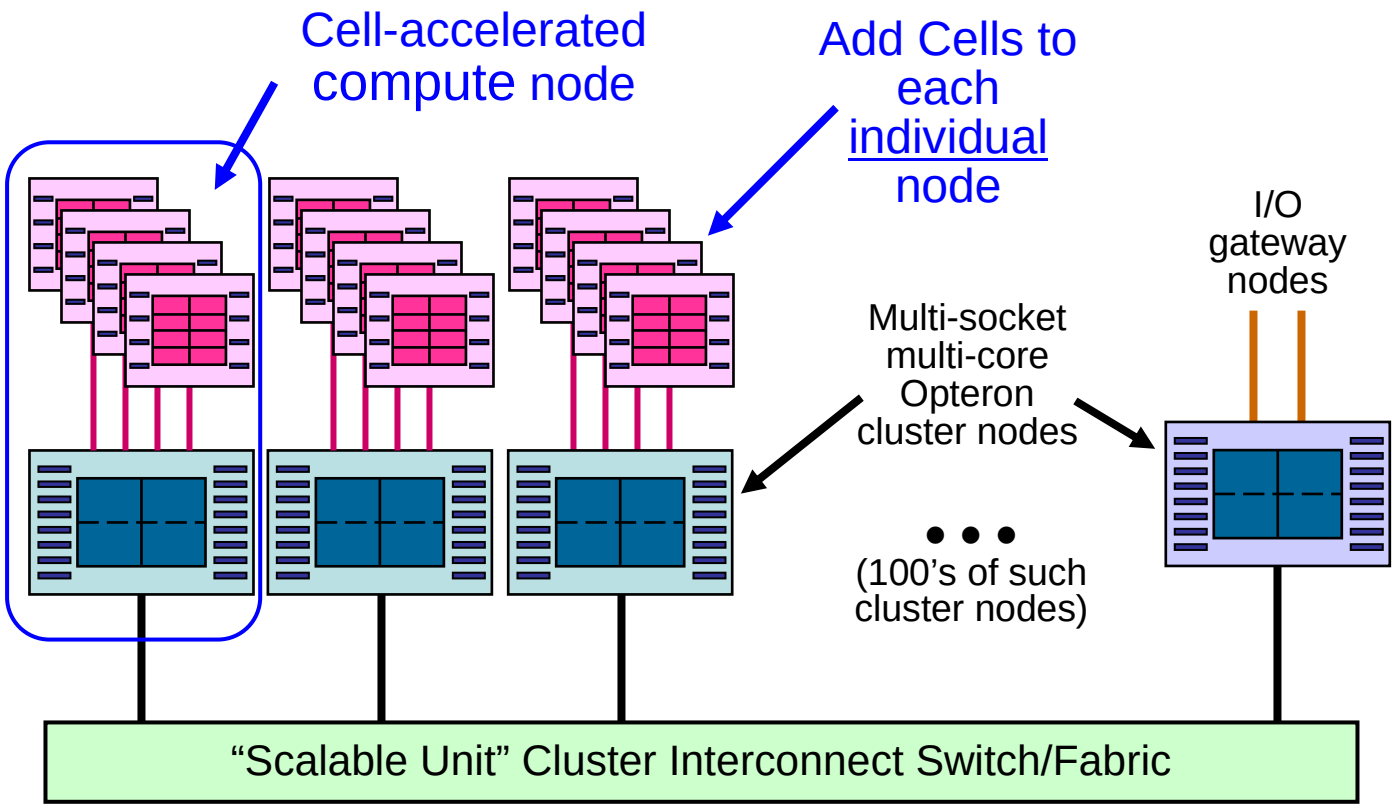
LANL has been looking at hybrid & petascale computing for some time



Roadrunner is on a different path to a petascale system

Roadrunner System Configuration

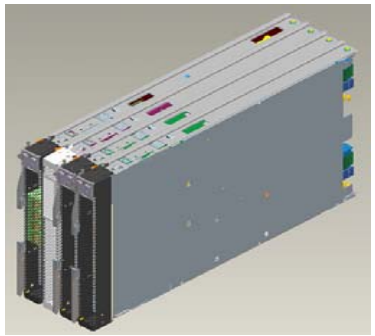
Roadrunner Phase 3 is Cell-accelerated, not a cluster of Cells

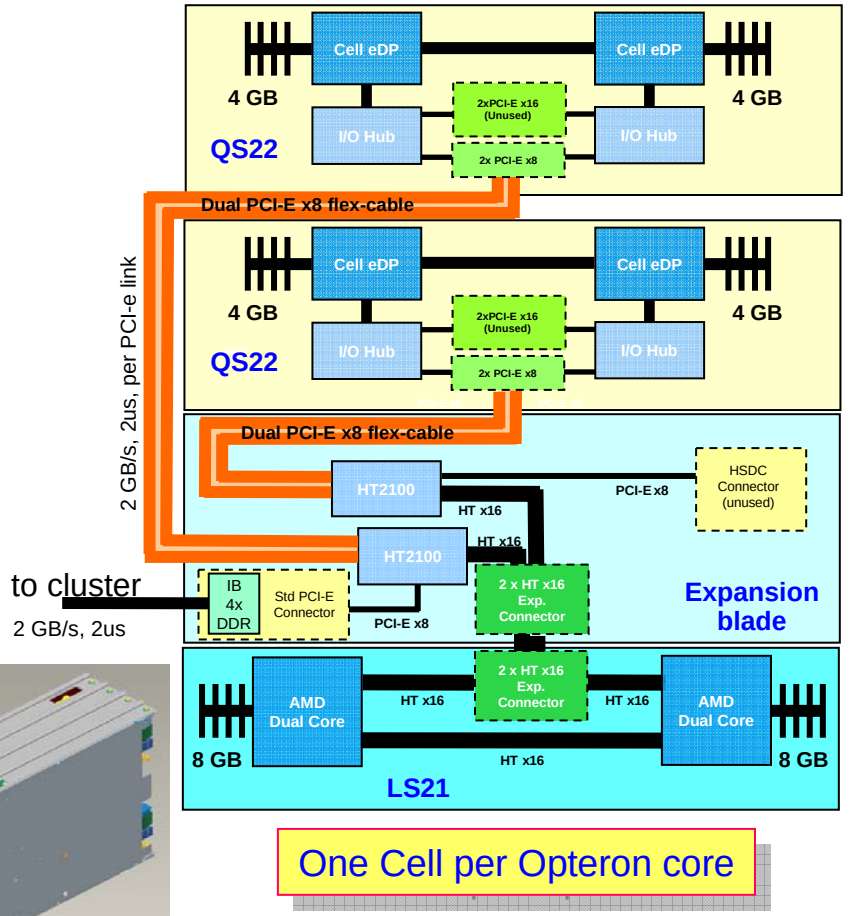


Node-attached Cells is what makes Roadrunner different!

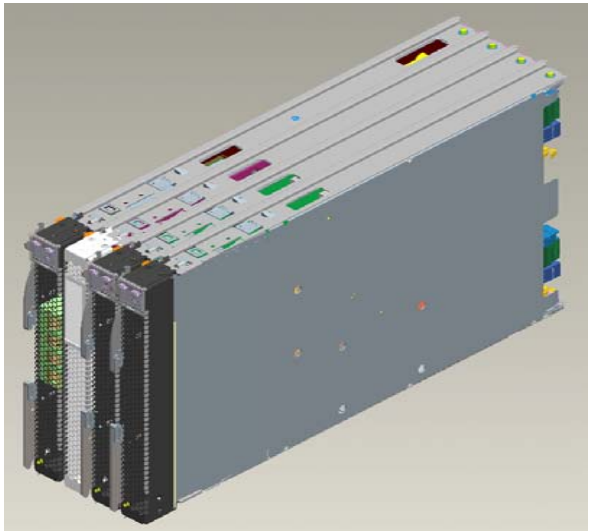
A Roadrunner TriBlade node integrates Cell and Opteron blades

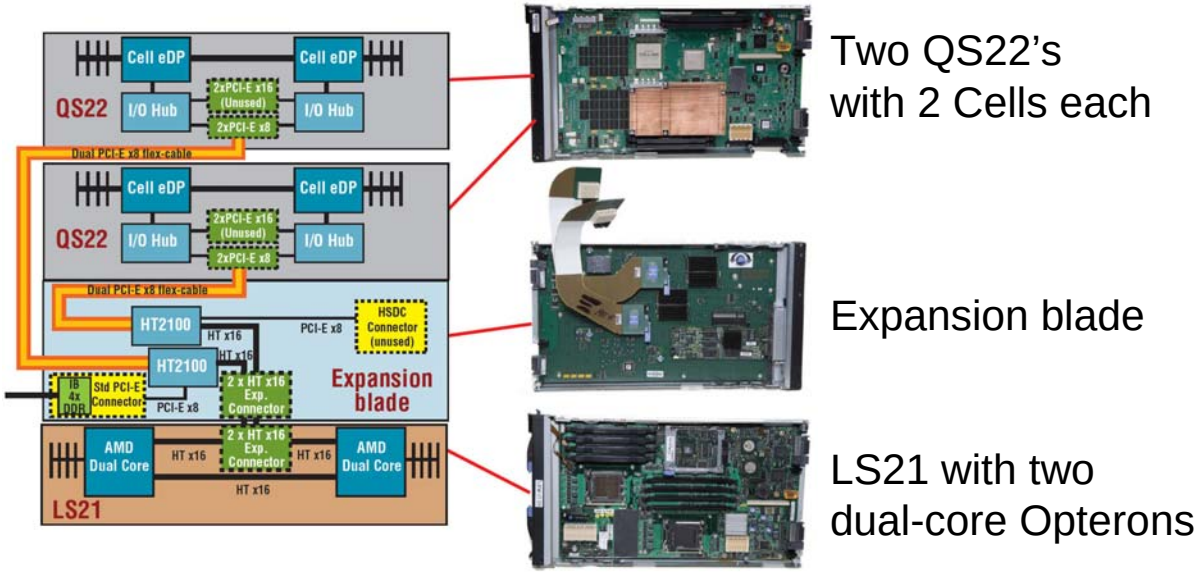
- QS22 is a future IBM Cell blade containing two new enhanced double-precision (eDP/PowerXCell™) Cell chips
- Expansion blade connects two QS22 via four PCI-e x8 links to LS21 & provides the node's ConnectX IB 4X DDR cluster attachment
- LS21 is an IBM dual-socket Opteron blade
- 4-wide IBM BladeCenter 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 Cell memory & 16 GB Opteron memory*



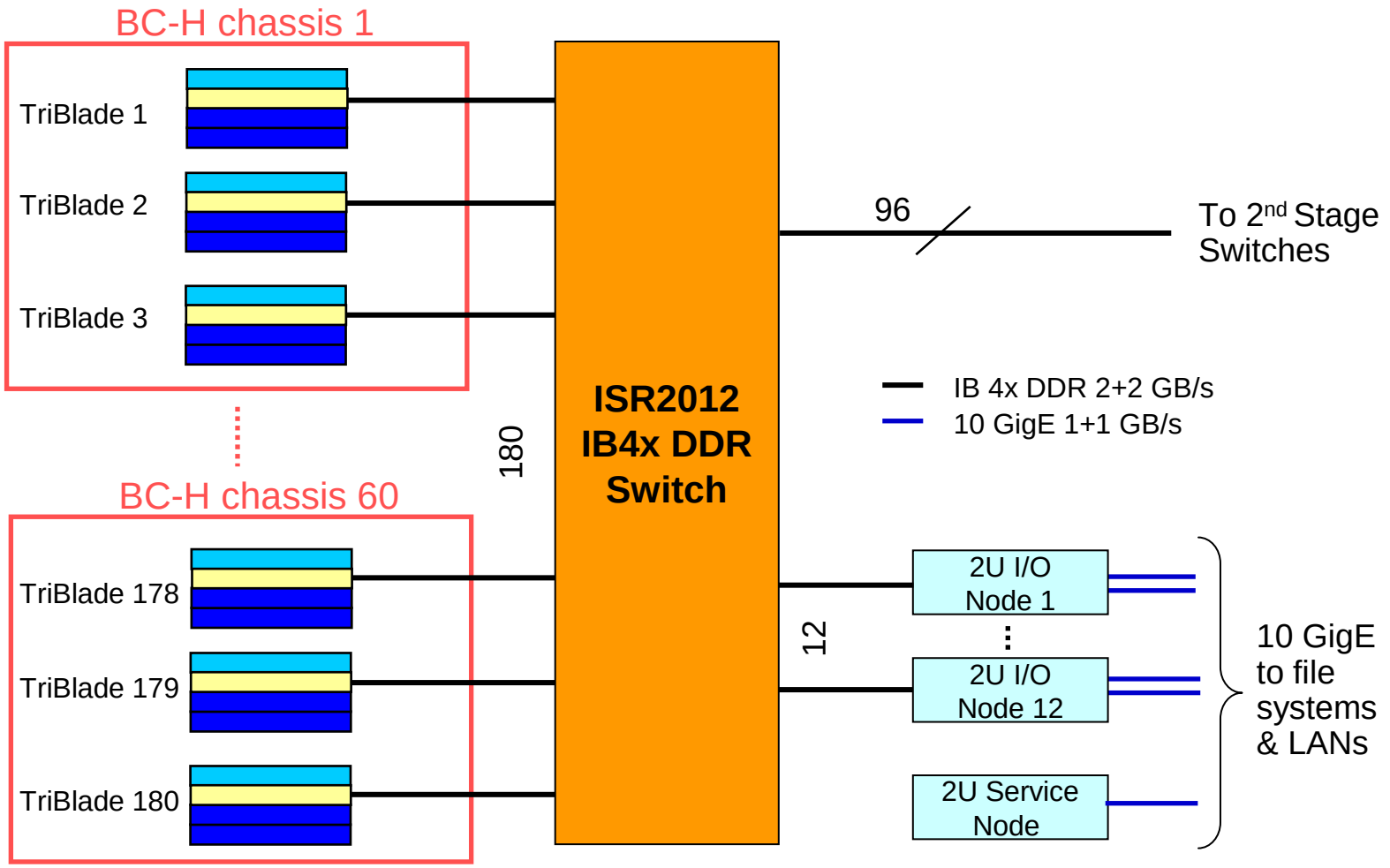


A Roadrunner TriBlade node integrates Cell and Opteron blades





A Connected Unit (CU) forms a building block



A Connected Unit (CU) is a powerful cluster

Connected Unit Specifications:

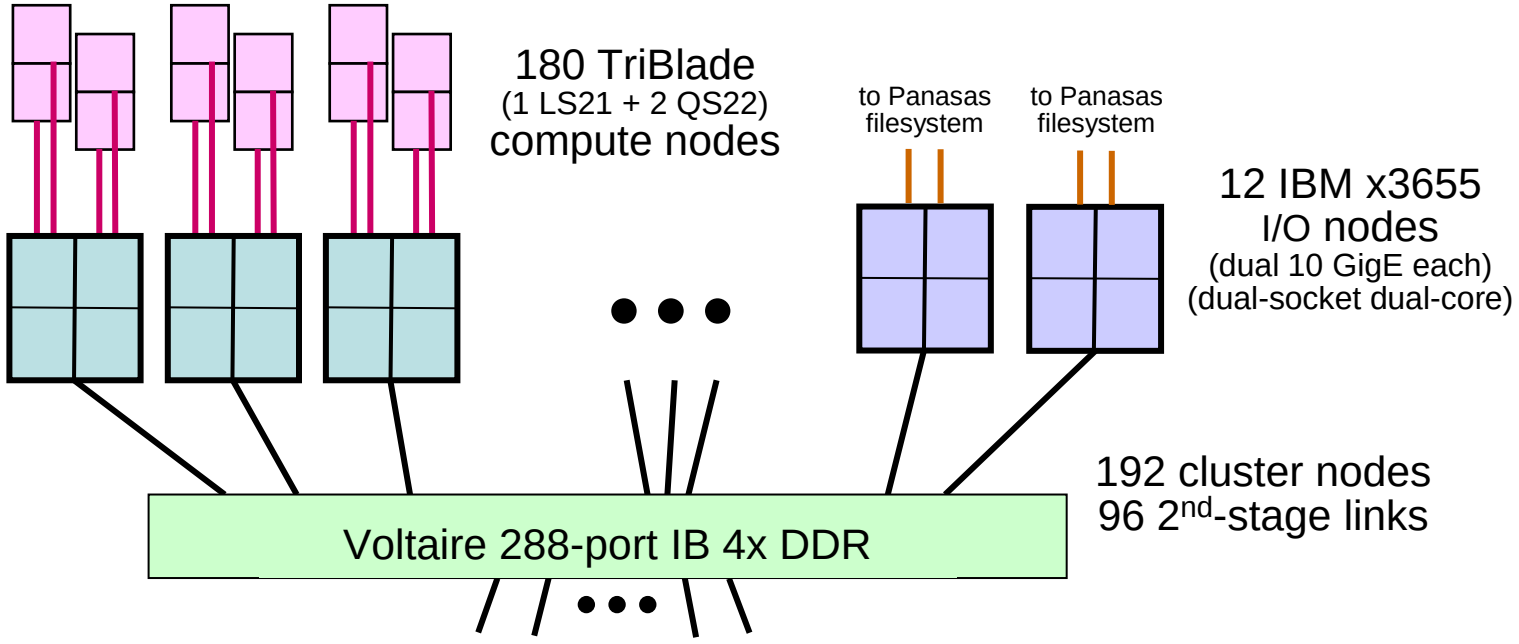
360 1.8 GHz dual-core Opterons
- 2.59 TF DP peak Opteron
- 2.88 TB Opteron memory

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

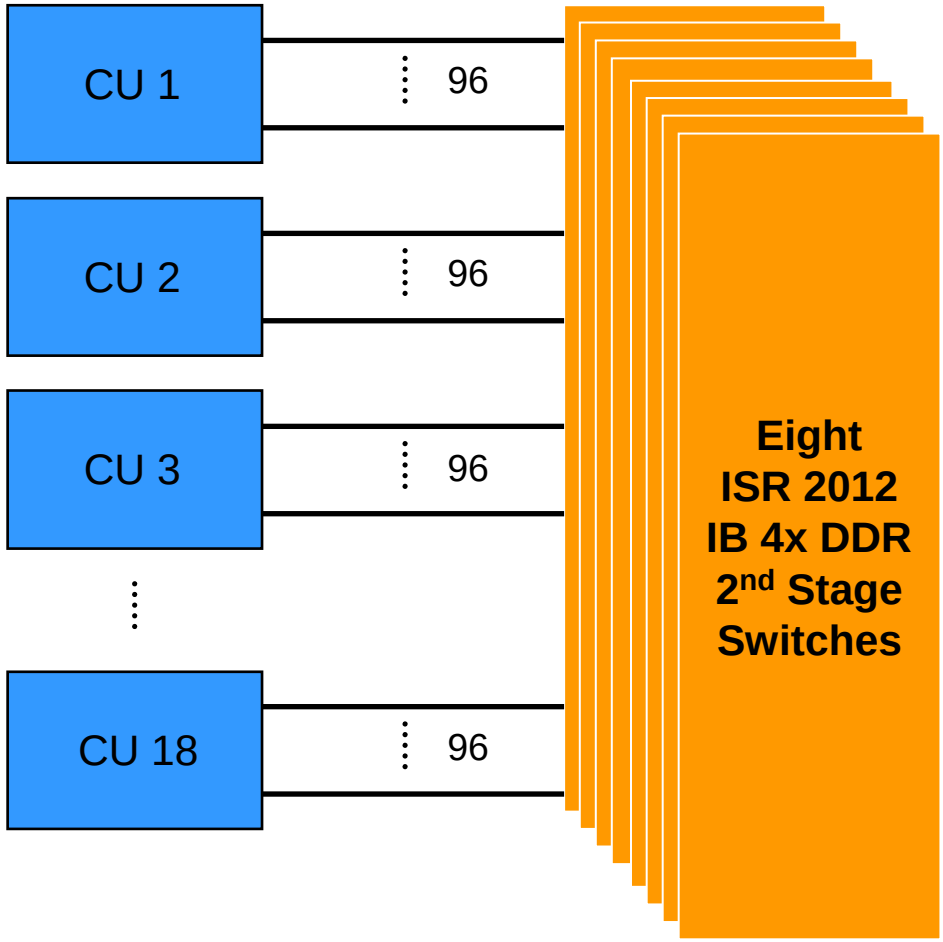
720 3.2 GHz Cell eDP chips
- 73.7 TF DP peak Cell eDP
- 2.88 TB Cell memory
- 18.4 TB/s Cell memory BW

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)



Now build a cluster-of-clusters…



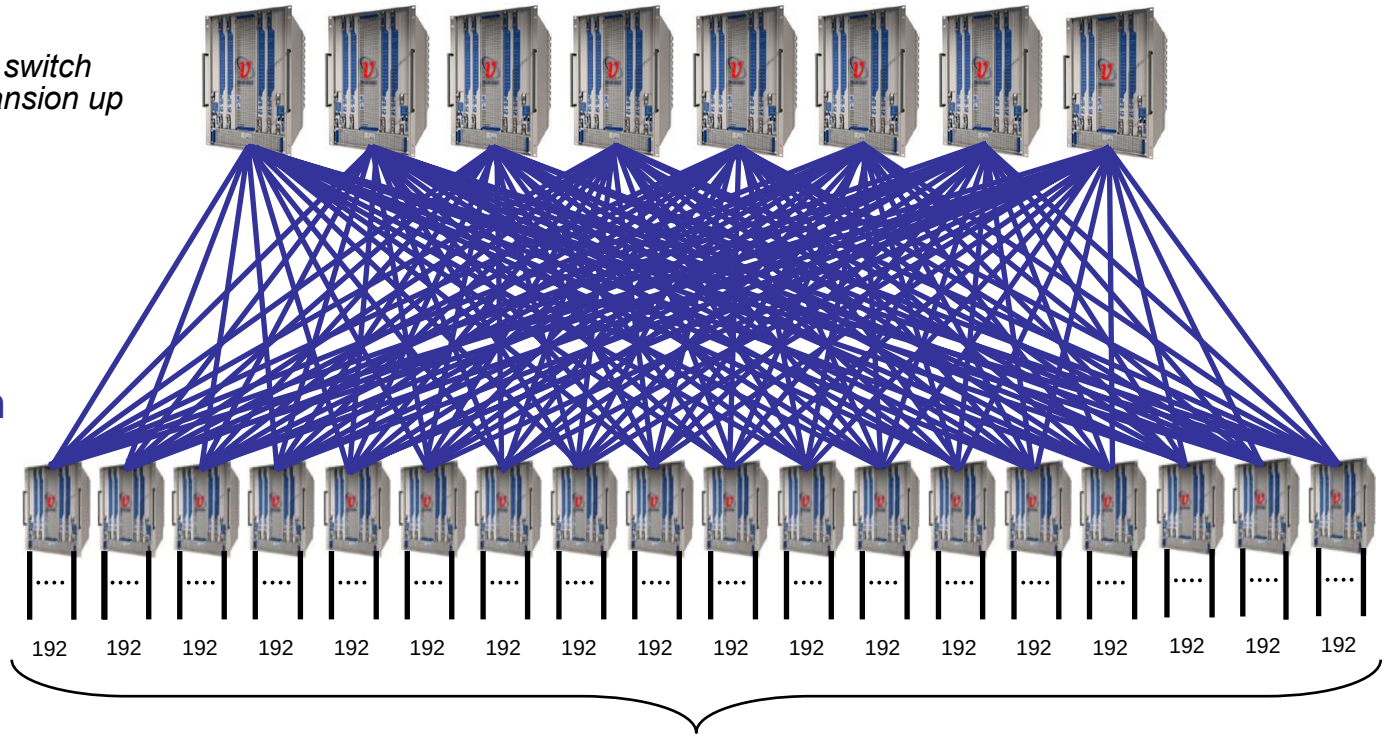
InfiniBand DDR 2–stage fat-tree interconnect

Eight 2nd–stage switches

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

12 uplinks per CU to each 2nd–stage switch provide half-bandwidth

Smaller systems (e.g. 4 CUs) by disabling links to other CUs for improved robustness during standup & stabilization



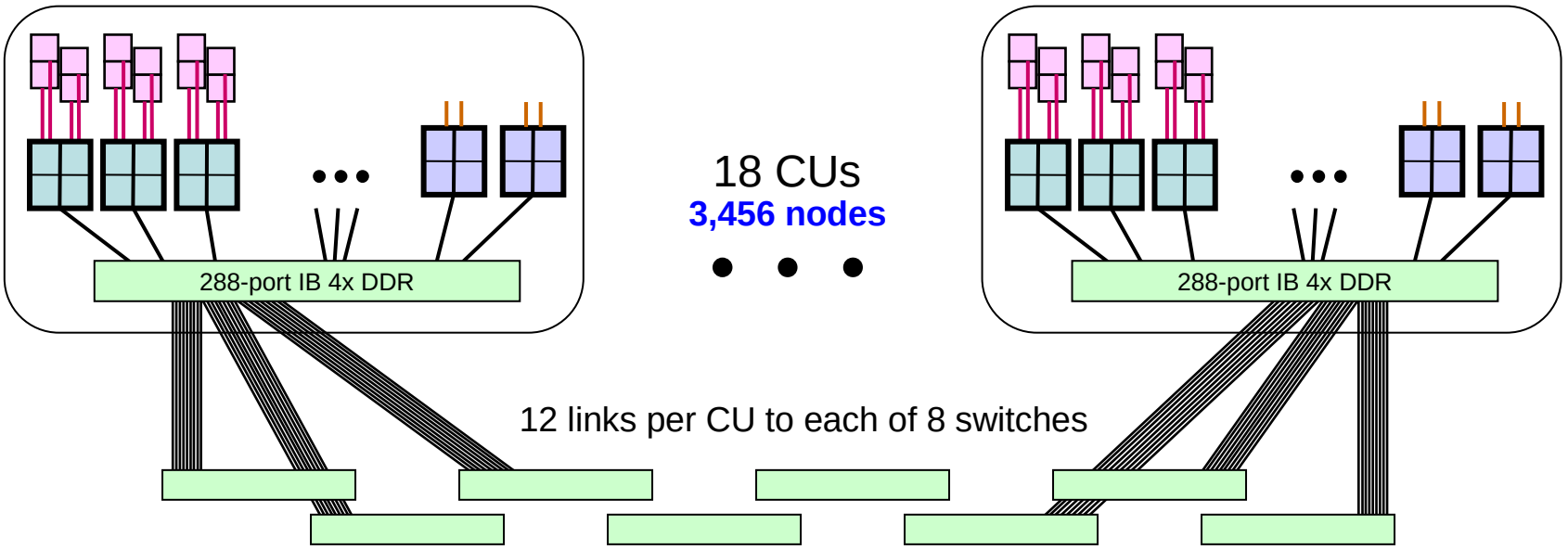
18 CUs with CU switches, 3456 IB nodes

Roadrunner is a hybrid petascale system of modest size delivered in 2008

Connected Unit cluster
180 compute nodes w/ Cells
12 x3655 I/O nodes

6,480 dual-core Opterons ⇒ 47 TF
12,960 Cell eDP chips ⇒ 1.3 PF

* I/O nodes not counted



Eight 2nd-stage 288-port IB 4X DDR switches

Roadrunner is a petascale system in 2008

Full Roadrunner Specifications:

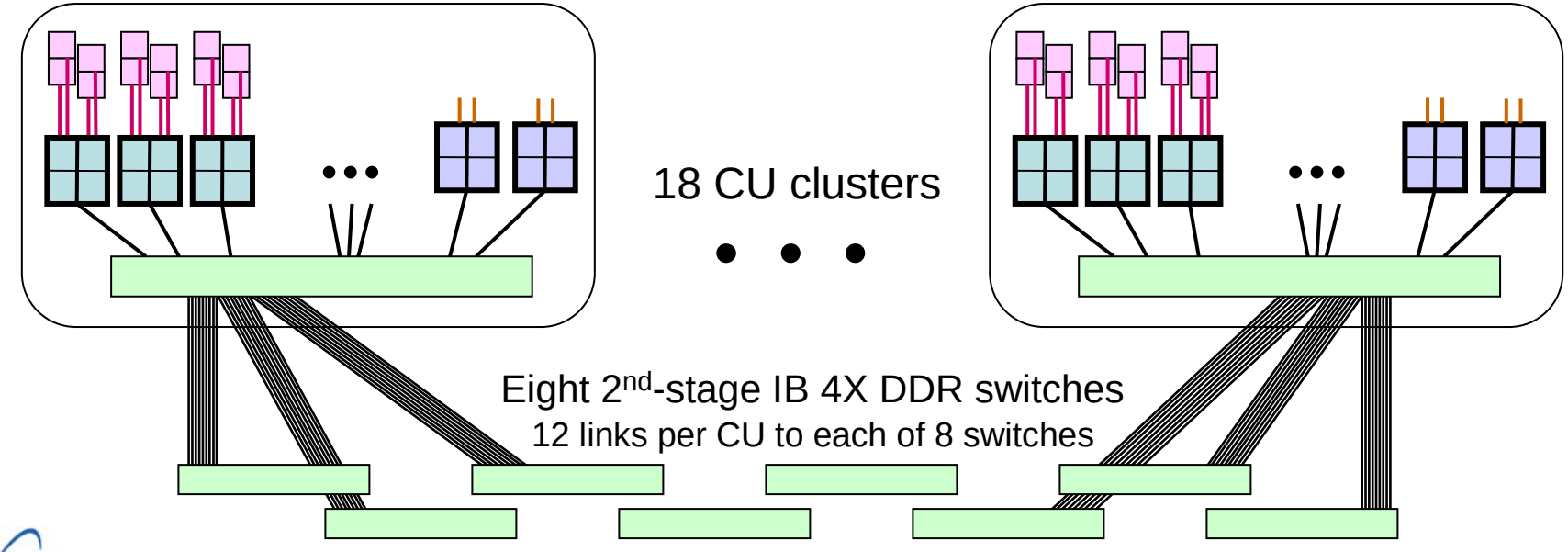
6,480 dual-core Opterons
- 46.7 TF DP peak Opteron
- 51.8 TB Opteron memory

432 dual-core Opterons
- in I/O nodes

12,960 Cell eDP chips
- *aka IBM PowerXCell™*
- 1.33 PF DP peak Cell eDP
- 2.65 PF SP peak Cell eDP
- 51.8 TB Cell memory
- 332 TB/s Cell memory BW

3,456 nodes on 2-stage IB 4X DDR
- 13.8 TB/s aggregate BW (bi-dir) (1st stage)
- 6.9 TB/s aggregate BW (bi-dir) (2nd stage)
- 3.5 TB/s bi-section BW (bi-dir) (2nd stage)

432 10 GigE I/O links on 216 I/O nodes
- 432 GB/s aggregate I/O BW (uni-dir)
- (IB limited)



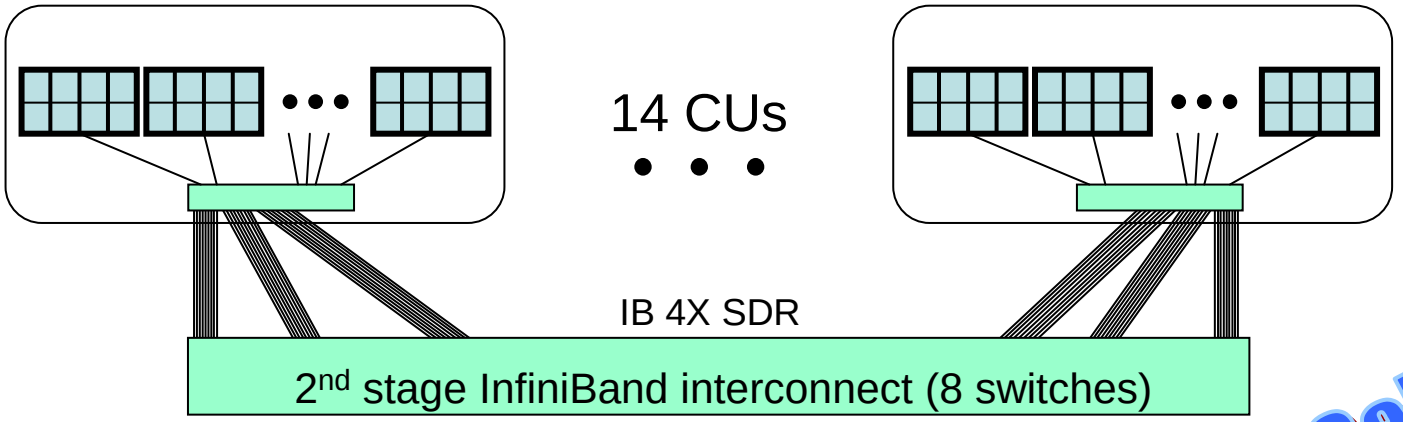
Roadrunner at a glance

- **Cluster of 18 Connected Units (CU)**
  - *6480 (+432) AMD dual-core Opterons*
  - *12,960 IBM Cell eDP accelerators*
  - *46.7 (+4.5) Teraflops peak (Opteron)*
  - *1.33 Petaflops peak (Cell eDP)*
  - *1PF sustained Linpack*
- **InfiniBand 4x DDR fabric**
  - *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*
    - 3.45 TB/s (bi-directional)
  - *Non-disruptive expansion to 24 CUs*
- **104 TB aggregate memory**
  - *52 TB Opteron*
  - *52 TB Cell*
- **216 GB/s sustained File System I/O:**
  - *216x2 10G Ethernets to Panasas*
- **Fedora Linux (RHEL possible)**
- **SDK for Multicore Acceleration**
  - *Cell compilers, libraries, tools*
- **xCAT Cluster Management**
  - *System-wide GigEnet network*
- **3.9 MW Power:**
  - *0.35 GF/Watt*
- **Area:**
  - *296 racks*
  - *5500 ft^2*

Early Roadrunner systems

Roadrunner Phase 1 "Base" systems: Redtail & Yellowrail

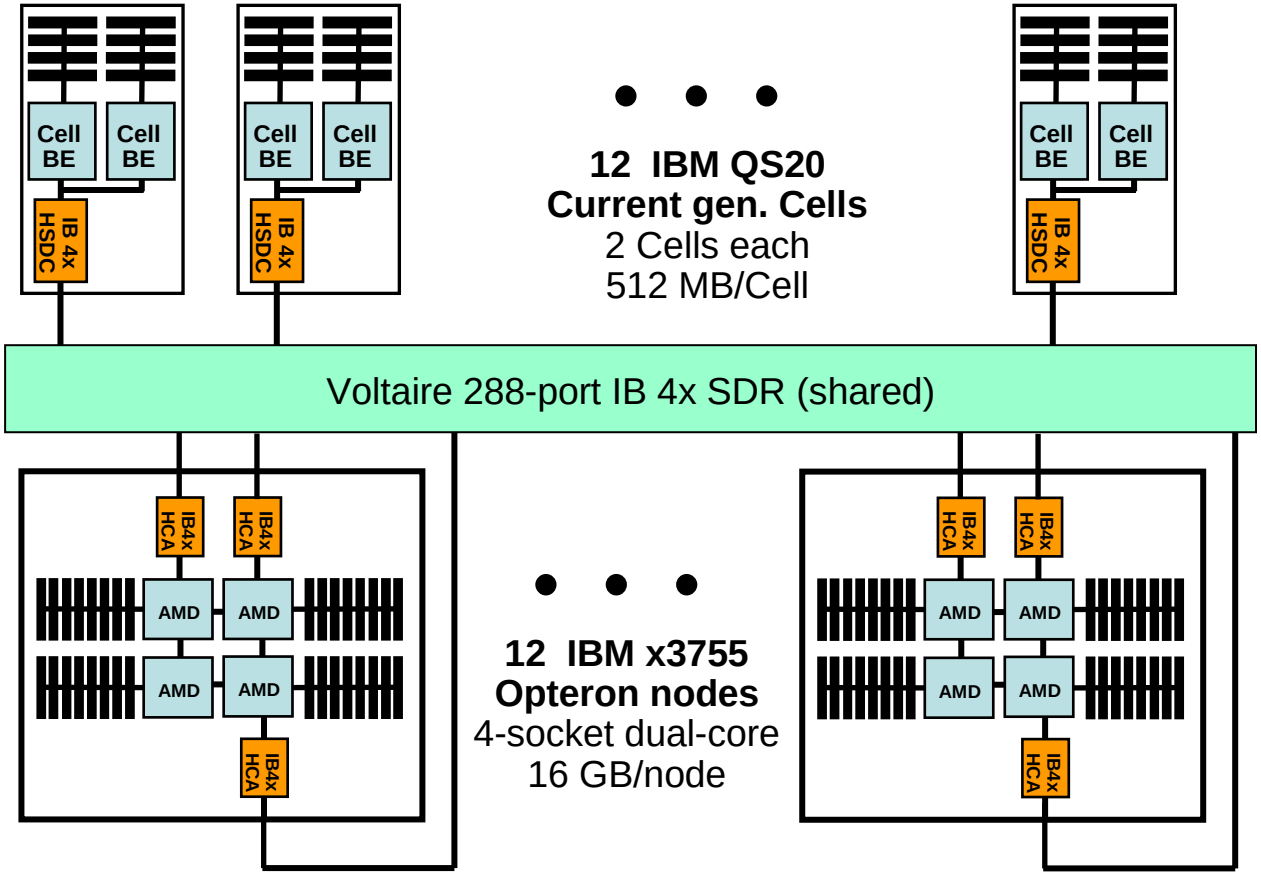
- Redtail is in the Secure in production
  - *Fourteen CUs of 8-way Opteron nodes (70 Teraflop/s)*
  - *144 4-socket dual-core 32 GB memory nodes (IBM x3755) per CU*
- Yellowrail is one CU in the Yellow



From earlier Roadrunner design

NO Cells

AAIS
Prototype for applications testing in 2007



Phase 2 prototype

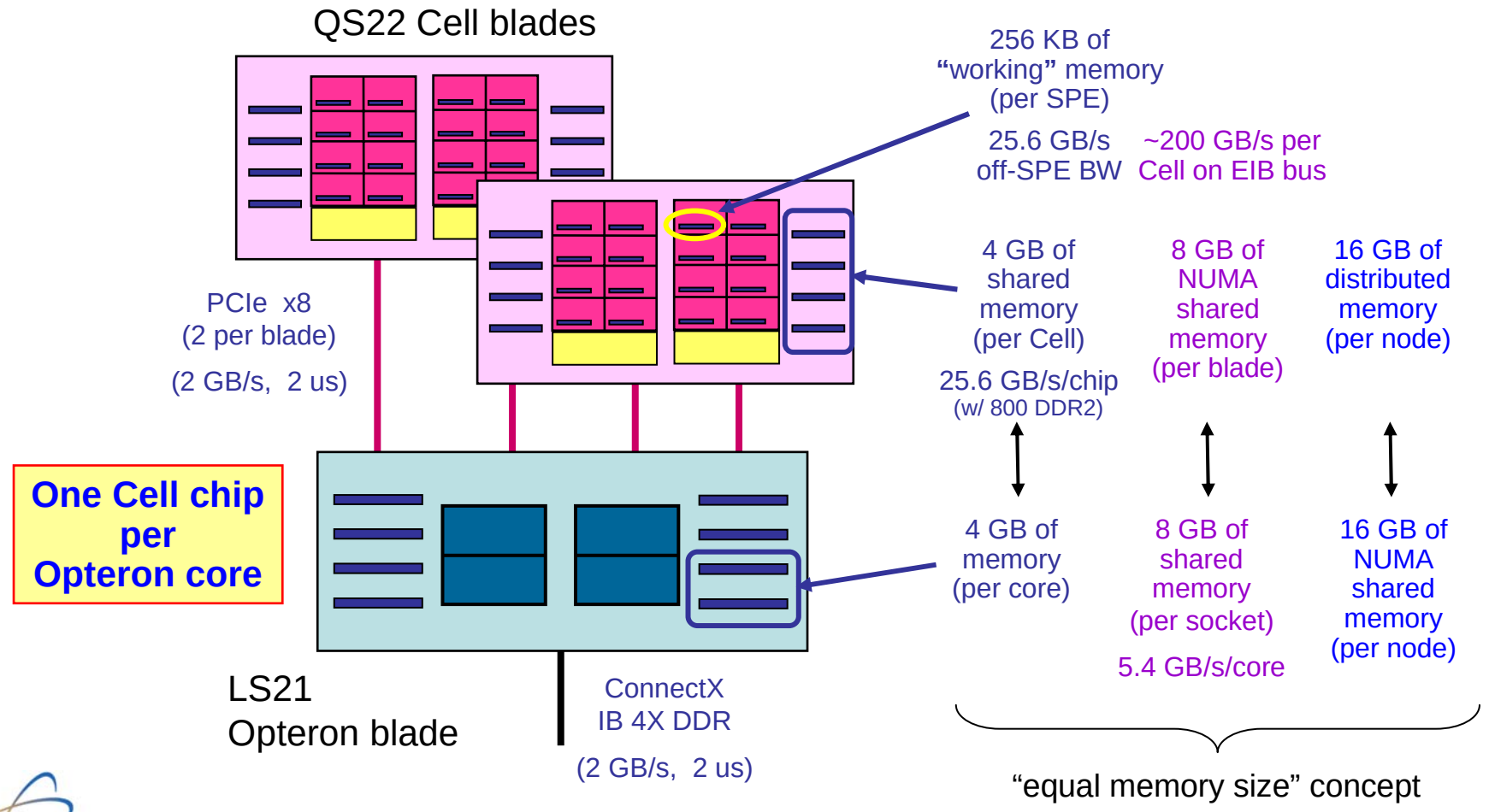
A dvanced
A rchitecture
I nitial
S ystem

aka. AAIS

(Operational January 2007)

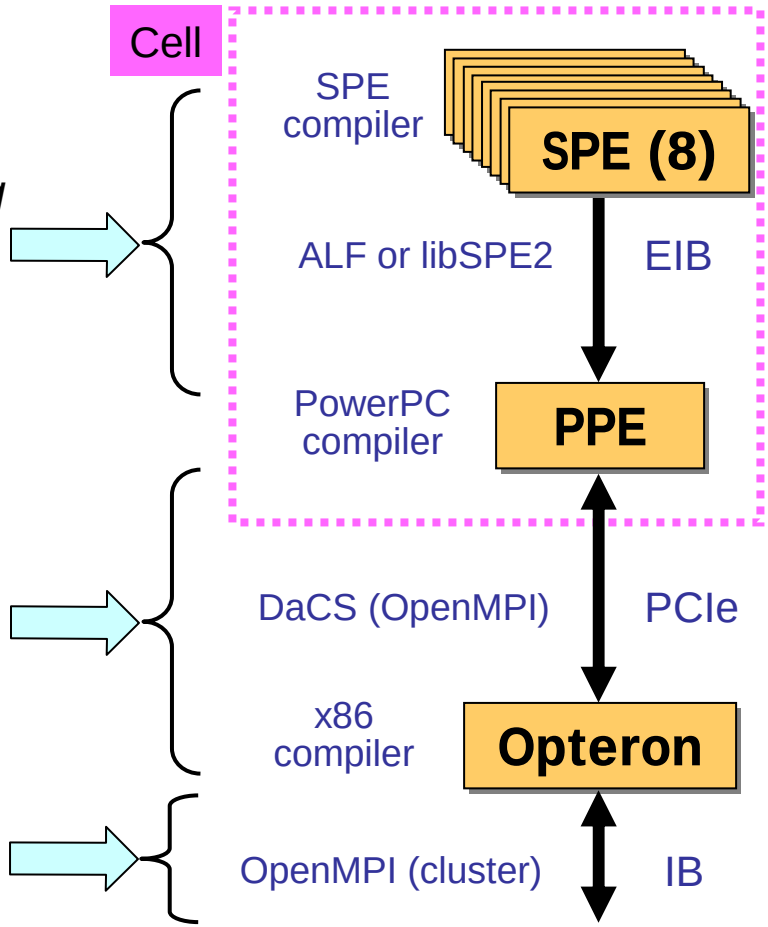
Programming Concepts

Roadrunner nodes have a memory hierarchy



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



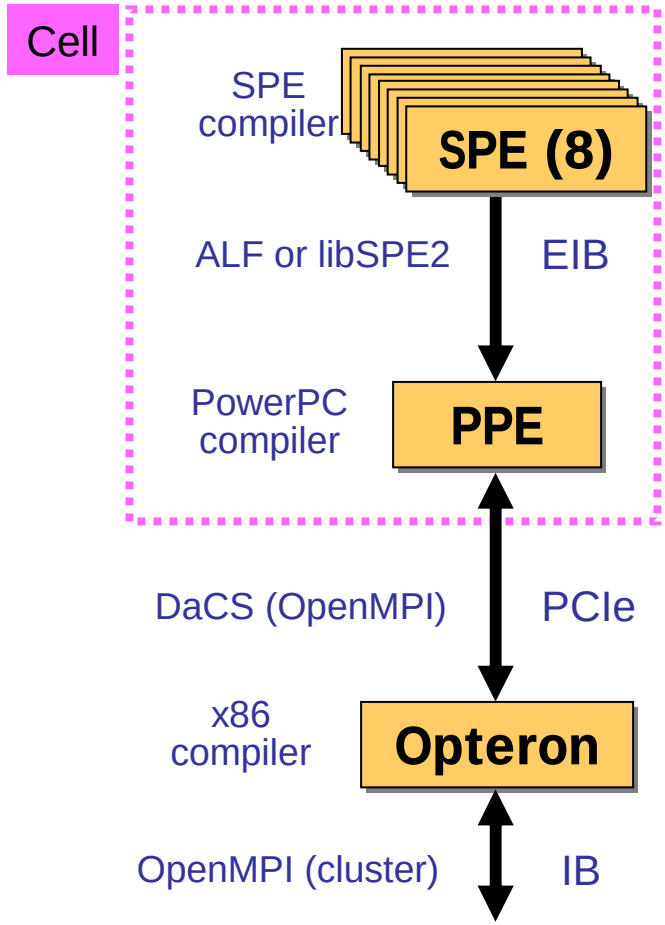
Three types of processors work together

Parallel-in-parallel

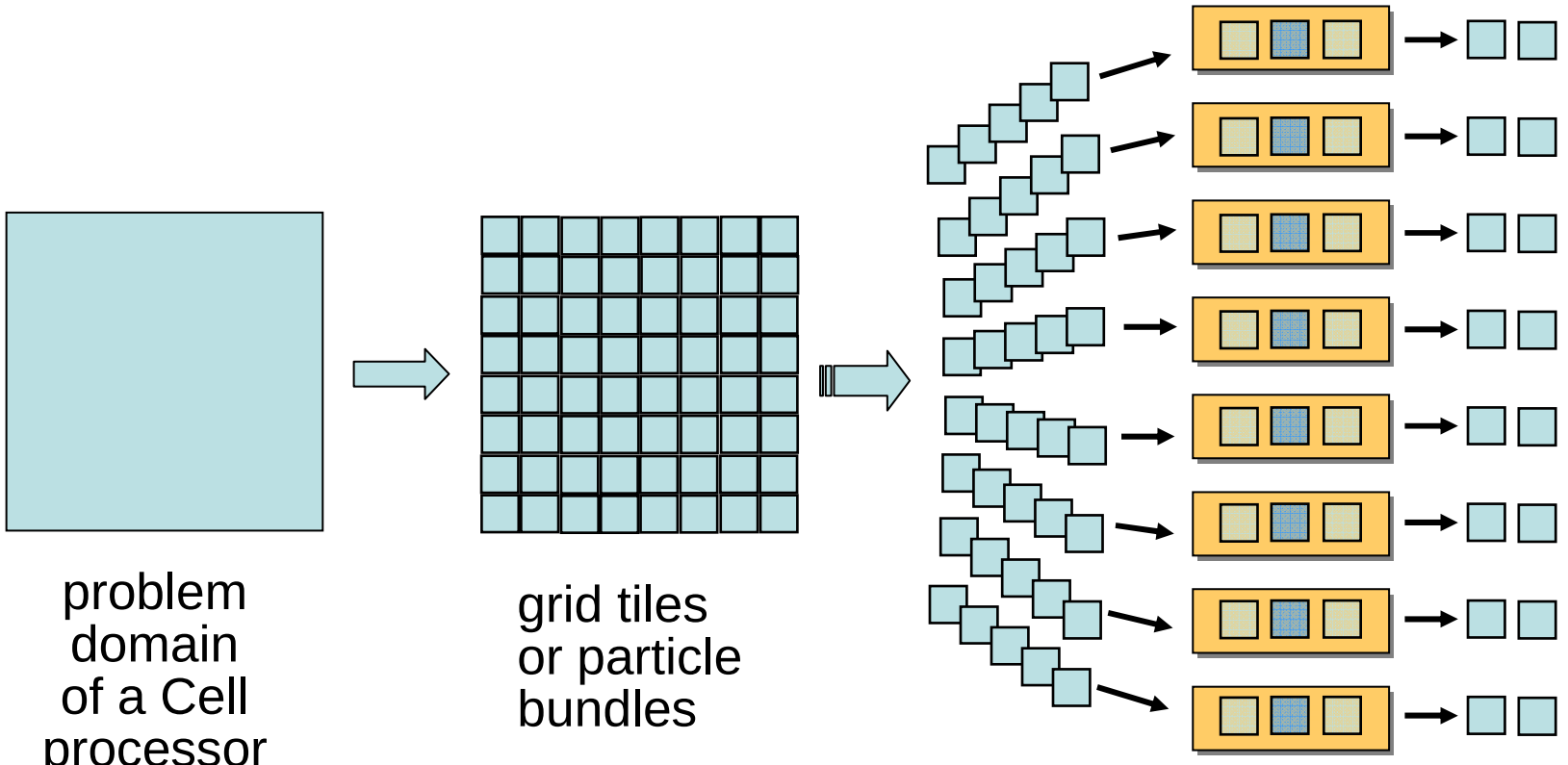
Questions to consider:

- Host-centric or Cell-centric
- Data transfers
- Cell local memory strategy

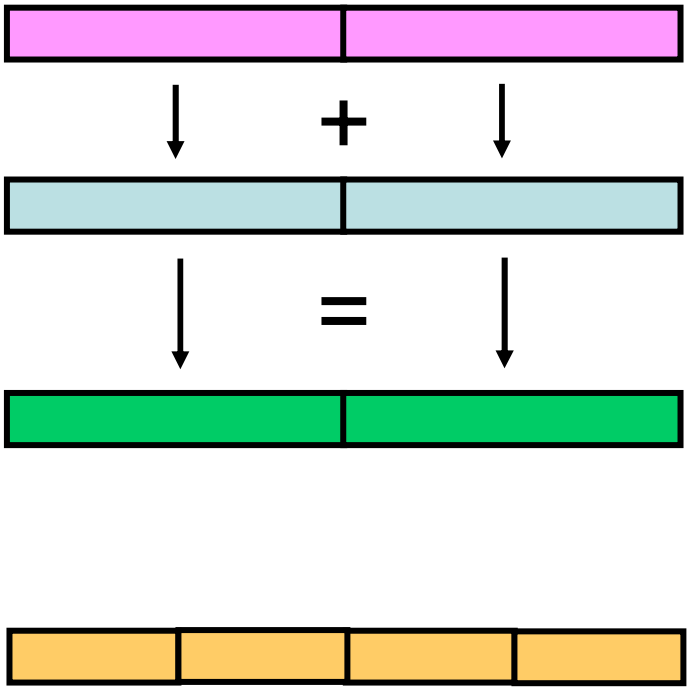
This can be done one algorithm at a time.



Break the work into a stream of pieces



Pick data structures & alignment to allow SIMD

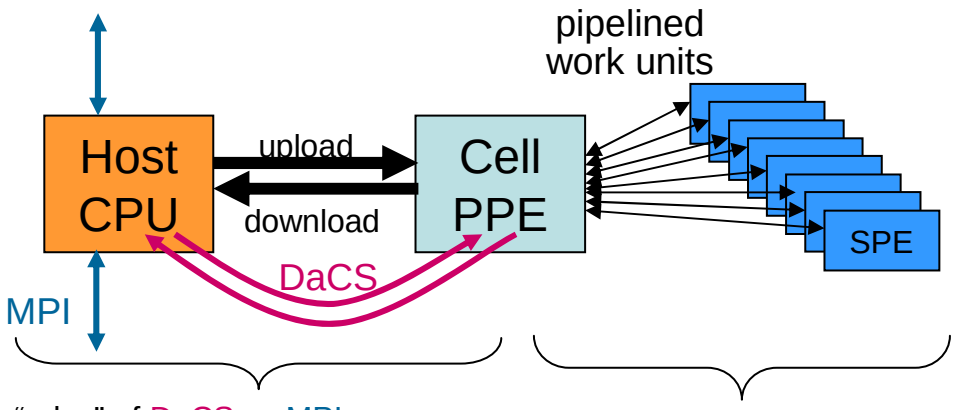


128 bits = 2 doubles
Work on aligned data
    c[i] = a[i] + b[i]

Cross aligned
operations
are really bad!
    c[i] = a[i] + a[i+1]

4 singles or integers
work similarly at
twice the performance

Put it all together



Compute & memory DMA transfers are overlapped in HW!

DMA Get (first)
Switch work buffers

DMA Get (prefetch)
DMA Wait (current get)
Compute
DMA Put (write 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);
```

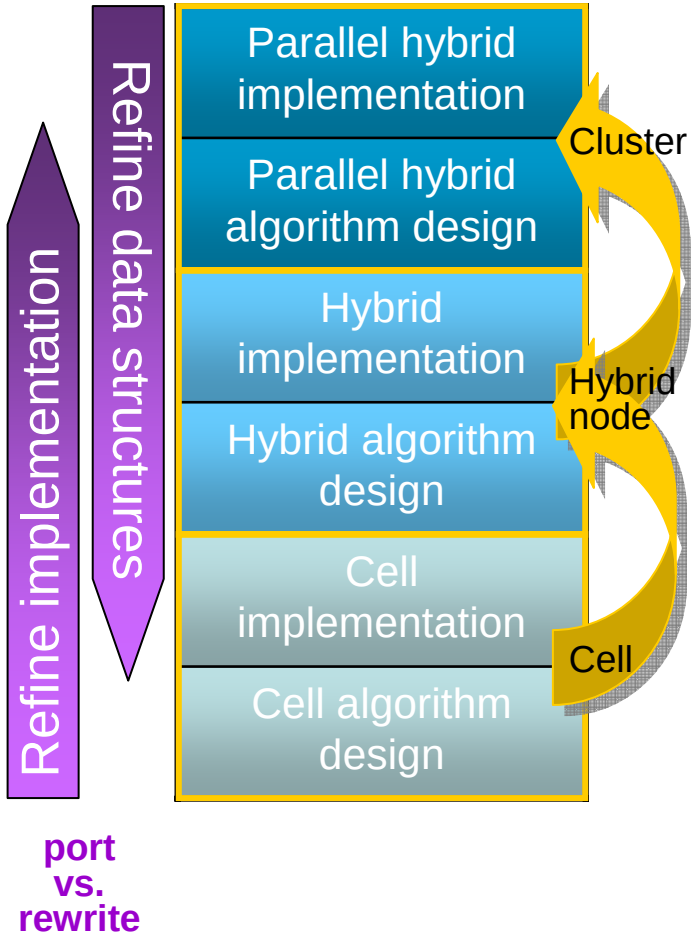
DMA Wait:
```
mfc_write_tag_mask(1<<tag);
mfc_read_tag_status_all();
```

Programming Approach is Tractable

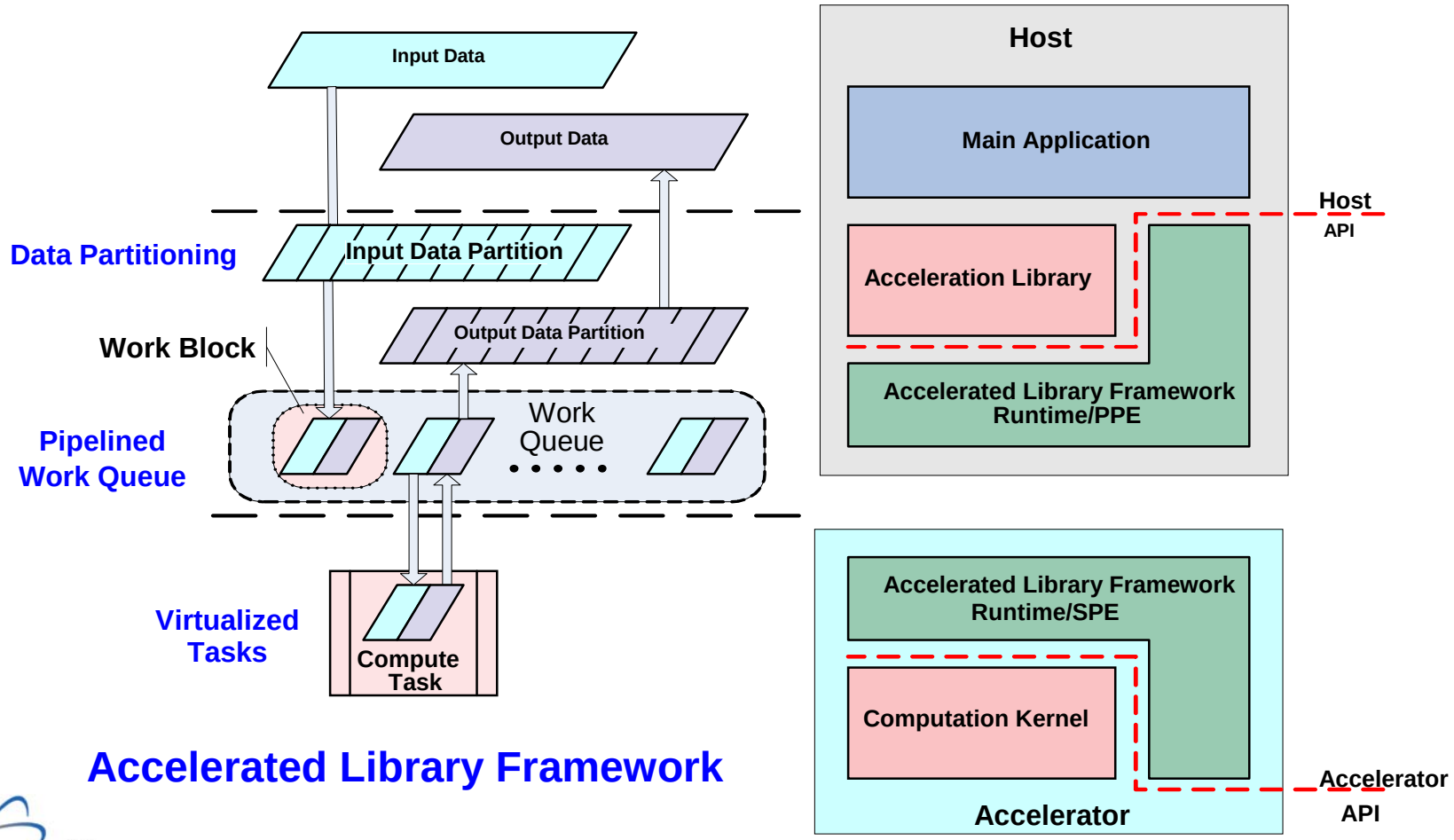
- Three levels of parallelism: node-to-node, within-node, within-Cell
- MPI for cluster-wide message passing still used between nodes
  - *Global Arrays, IPC, UPC, Global Address Space (GAS) languages, etc. also remain possible choices*
  - *Additional parallelism can be introduced within the node ("divide & conquer")*
    - Roadrunner only has ~3000 4-way compute nodes so it may not be required
- Split off large-grain computationally intense portions of code for Cell acceleration within a node process
  - *This is equivalent to function offload*
    - One Cell per one Opteron core
    - Opteron would typically block, but could do concurrent work
  - *Additional fine-grained parallelism introduced within the Cell itself*
    - Create many-way parallel pipelined work units for SPMD on the SPEs
      - *MPMD, RPCs, streaming, etc. are also possible*
  - *Embedded MPI communications are possible via "relay" approach*
  - *Consistent with heterogeneous chips future trends*
- Considerable flexibility and opportunities exist

A few key thoughts for Roadrunner codes

- *3 cooperating programs*
  - some could simply be “relays”
- *Cells can’t talk to other Cells, only it’s host Opteron*
- *Where do I put “main” logic & memory*
  - *Opteron or Cell PPE?*
- *Break all work into streams of fixed-sized tiles or chunks for SPEs*
- *Use DMA transfers and reusable work buffers in SPE local memory to keep streams flowing*

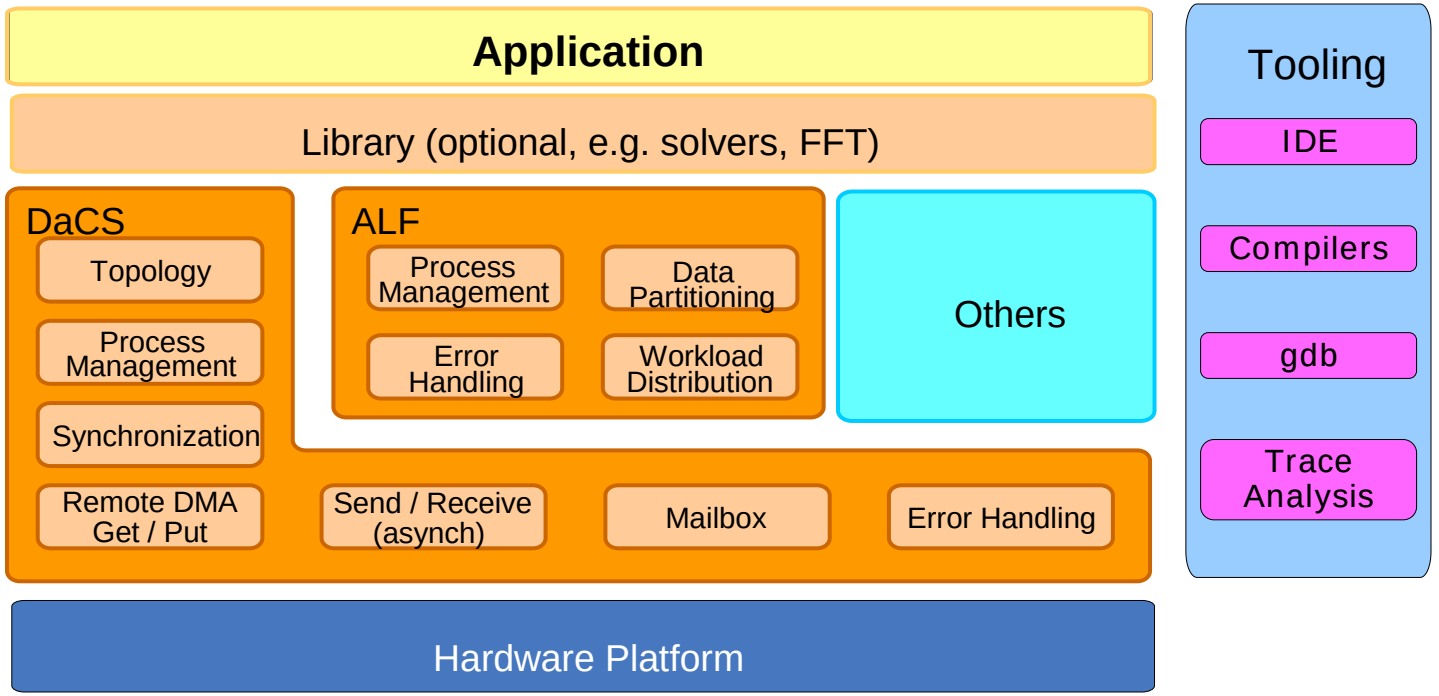


IBM-provided ALF is a simple work-queue approach



Accelerated Library Framework

ALF & DaCS: Broader than Cell & Roadrunner



- Designed by IBM & LANL to be HW agnostic
  - multicore/GPU/Cell, interconnect, even possibly cluster-wide
  - desire technical community participation to extend range

Roadrunner embodies many key architectural trends, each in moderation.

- Roadrunner has: multicore, short-vector SIMD, threads, heterogeneous instruction sets, local stores instead of caches, explicit DMA, on-chip CPU/memory networks, remote accelerators and cluster computing.
- You can use any of these features as needed, without needing to go to extremes in any one of them.
- Roadrunner's scale and flexibility makes it an ideal base from which to explore the changing landscape of HPC
  - *But it also provides immediate benefits!*

More information

The Roadrunner Technical Seminar Series

- **March 13: "Roadrunner Platform Overview," Ken Koch, CCS-DO.**
  Introduction to the final Roadrunner system hardware and nuances. Intended for people unfamiliar with the system and underlying technology.
- **March 18: "Overview of Applications, Results, and Programming," John Turner, CCS-2**
  Overview of applications, technical results, and programming approaches. Intended to prepare the audience for more detailed application talks to follow.
- **March 19: "Overview of Modeling, Performance, and Results," Darren Kerbyson, CCS-1**
  Overview of modeling and prediction approaches for Roadrunner, providing technical details as to how performance modeling work is done and its impact on systems and applications.
- **April 10: "Application 1: SPaSM," Sriram Swaminarayan, CCS-2**
  Discussion of the SPaSM application and what was done to put it on Roadrunner, what results were obtained, what was learned, and what is next.
- **April 22: "Application 2: VPIC," Ben Bergen, CCS-2**
  Discussion of the VPIC application and what was done to put it on Roadrunner, what results were obtained, what was learned, and what is next.
- **April 23: "Application 3: SWEEP3D," Mike Lang, CCS-1**
  Discussion of the SWEEP3D application kernel and what was done to put it on Roadrunner, what results were obtained, what was learned, and what is next.
- **April 24: "Application 4: Milagro I," Tim Kelley, CCS-2**
  Discussion of the Milagro application and one approach to put it on Roadrunner, what results were obtained, what was learned, and what is next.
- **May 6:"Application 5: Milagro II," Paul Henning, CCS-2**
  Discussion of a different approach to put Milagro on Roadrunner, what results were obtained, what was learned, and what is next.
- **May 8: "Application 6: DNS," Jamal Mohd-Yusof, CCS-2**
  Why Direct Numerical Simulation is important and what was done to put it on Roadrunner, what results were obtained, what was learned, and what was next.
- **May 29: "Panel Discussion: Hybrid Computing Programming Models," moderated by John Turner, CCS-2.**
  Panel participants TBD.
  Discussion of various programming approaches on Roadrunner in particular and hybrid computing in general, tools, and future activities.
- **June 3: "Panel Discussion: Future Platforms," moderated by Adolfy Hoisie, CCS-1.**
  Panel participants TBD.
  Discussion of future computing platforms and how they will be similar to, and differ from, Roadrunner.

LANL Roadrunner web sites

More information is available at:

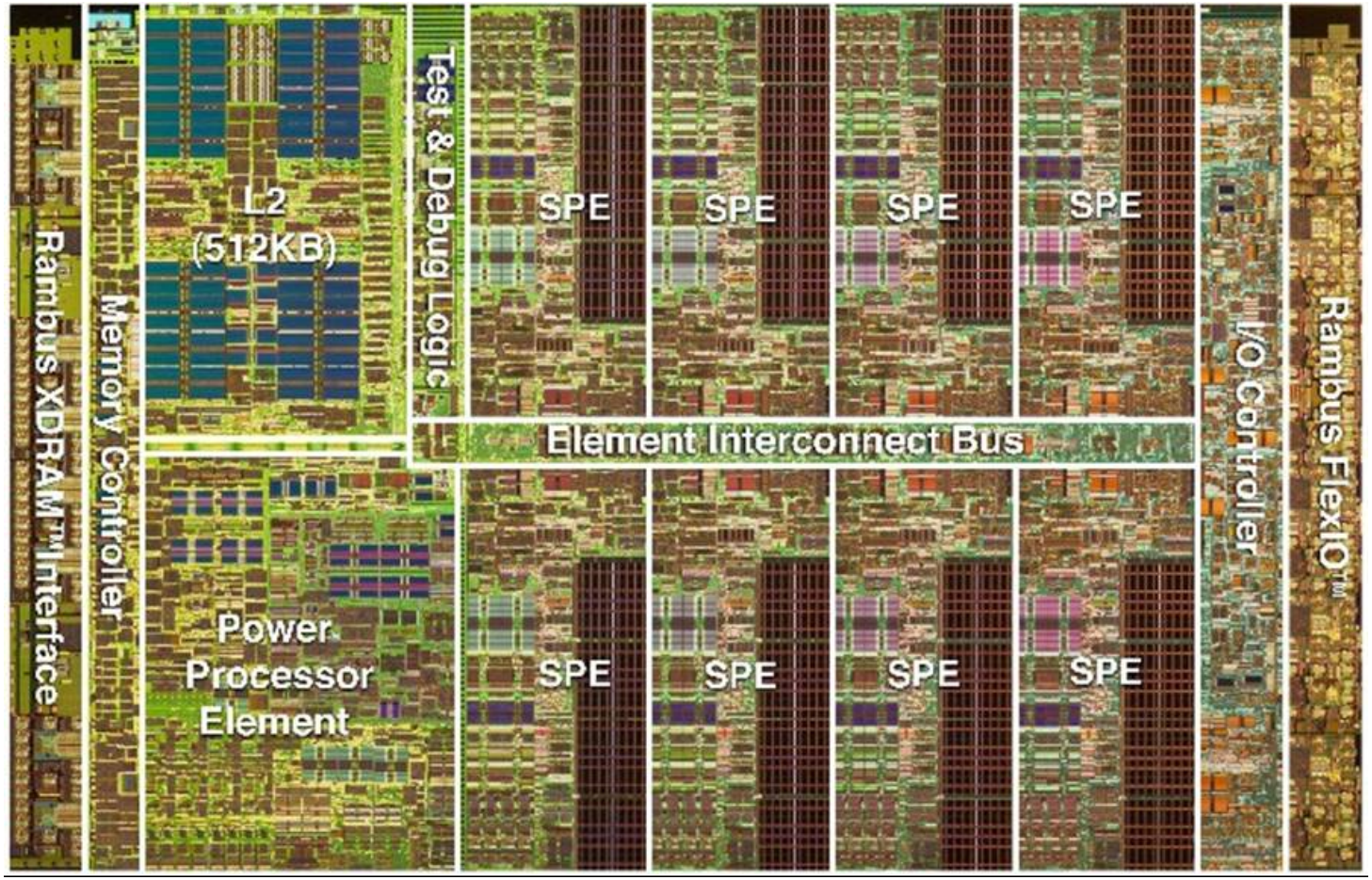
http://www.lanl.gov/roadrunner/ *(Internet)*

http://rralgs.lanl.gov/portal *(LANL Yellow)*

Roadrunner Architecture
Roadrunner Applications efforts
Roadrunner Programming
Other Roadrunner talks
Computing Trends
Related Internet links

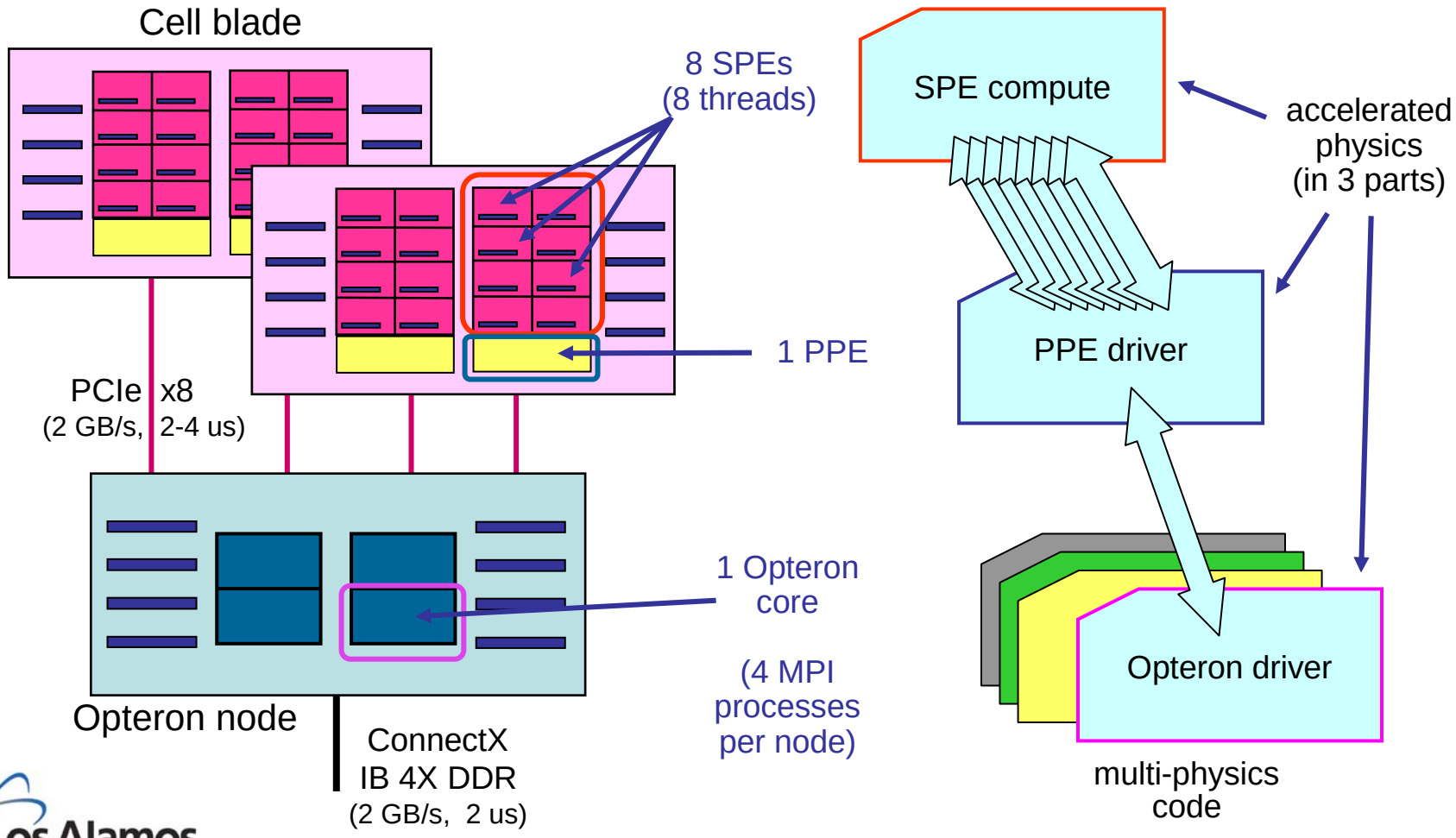
Extra slides

Cell Broadband Engine



Heterogeneous: 1PPE + 8 SPEs

One approach to using Roadrunner's processor hierarchy.



Before Roadrunner …

- Floating Point Systems FPS Array Processors (AP-120B, FPS-164/264) (circa 1976-1982)
  - http://en.wikipedia.org/wiki/Floating_Point_Systems
- Deep Blue for chess (IBM SP-2: 30 RS6K + 480 chess chips) (circa 1997)
  - http://en.wikipedia.org/wiki/Deep_Blue
- Grape-6 for stellar dynamics w/ custom chips) (circa 2000-2004)
  - http://grape.astron.s.u-tokyo.ac.jp/~makino/grape6.html
- Various FPGA supercomputers from system vendors:
  - *SRC-6 (w/ MAP)*
  - *Cray XD1 (w/ Application Acceleration)*
  - *SGI Altix (w/ RASC)*
- Titech TSUBAME (w/ some Clearspeed) (2006)
  - http://www.gsic.titech.ac.jp/English/Publication/pressrelease.html.en
- RIKEN MDGrape-3 “Protein Explorer” (w/ custom chips) (2006)
  - http://mdgrape.gsc.riken.jp/modules/tinyd0/index.php
- Terra Soft’s Cell E.coli/Amoeba PS3 Cluster (cluster of 1U PlayStation 3 development systems) (2007)
  - http://www.hpcwire.com/hpc/967146.html