Roadrunner Goes Beyond MPI

Next Generation Scalable Applications: When MPI-only is not enough

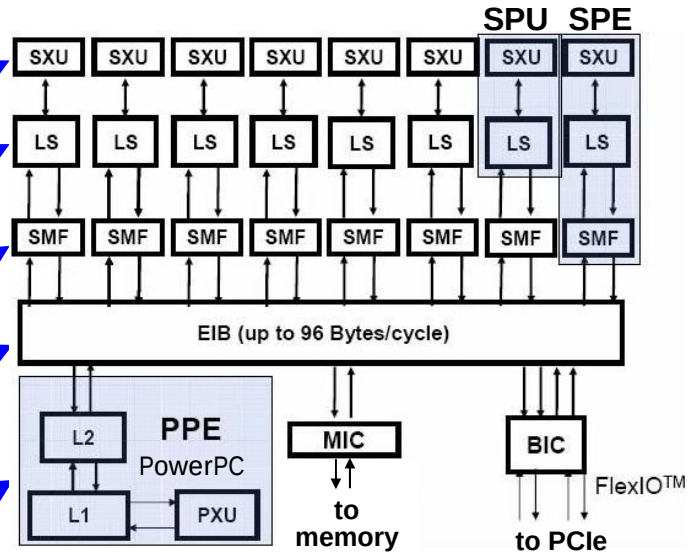
June 3-5, 2008

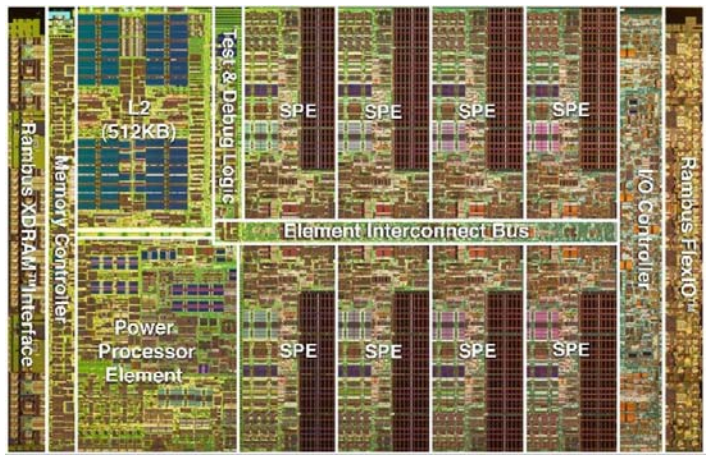
Ken Koch

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

Roadrunner uses a new enhanced double-precision variant of the Cell processor chip

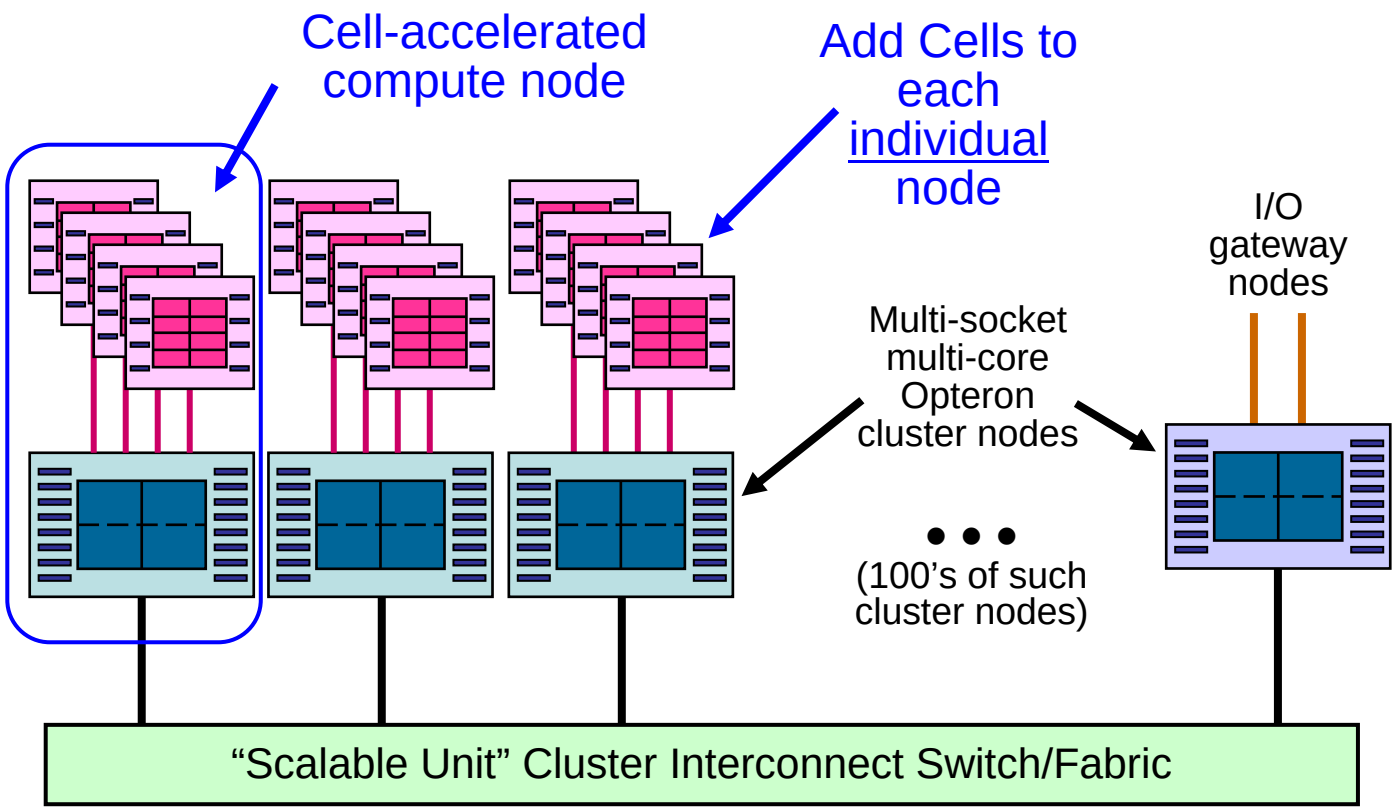
- Cell Broadband Engine (CBE*) developed by Sony-Toshiba-IBM
  - Used in Sony PlayStation 3
  - 8+1 hybrid processor chip
- 8 Synergistic Processing Elements (SPEs)
  - 128-bit SIMD-vector engines
  - 256 kB local memory (LS = Local Store)
  - Direct Memory Access (DMA) engine (25.6 GB/s each)
  - On-chip interconnect (EIB)
  - Run SPE-code as POSIX threads (SPMD, MPMD, streaming)
- PowerPC PPE runs Linux OS
- **New** Enhanced-DP Cell chip: (IBM PowerXCell 8i)
  - 102.4 GF/s DP (204.8 GF/s SP still)
  - 4-8 GB @ 25.6 GB/s DDR2 memory





* trademark of Sony Computer Entertainment, Inc.

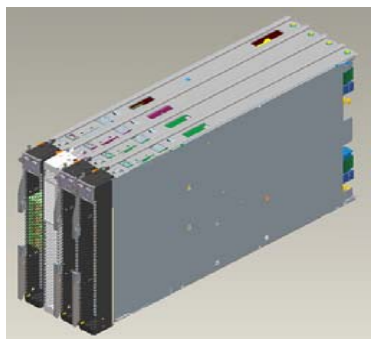
Roadrunner Phase 3 is Cell-accelerated, not a cluster of Cells

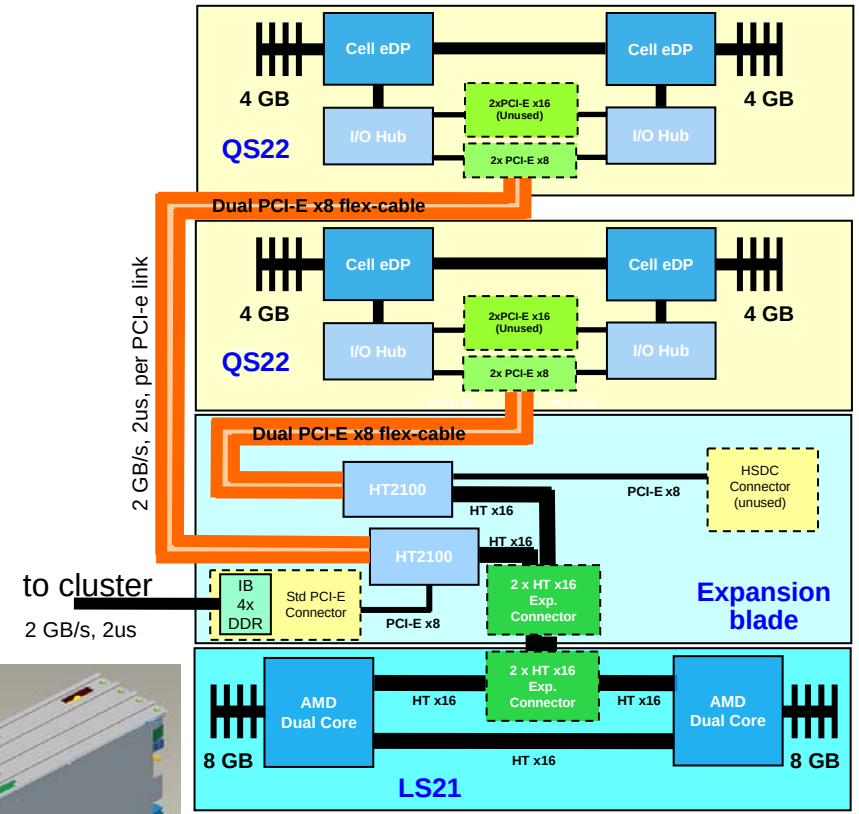


Node-attached Cells is what makes Roadrunner different!

A Roadrunner TriBlade node integrates Cell and Opteron blades

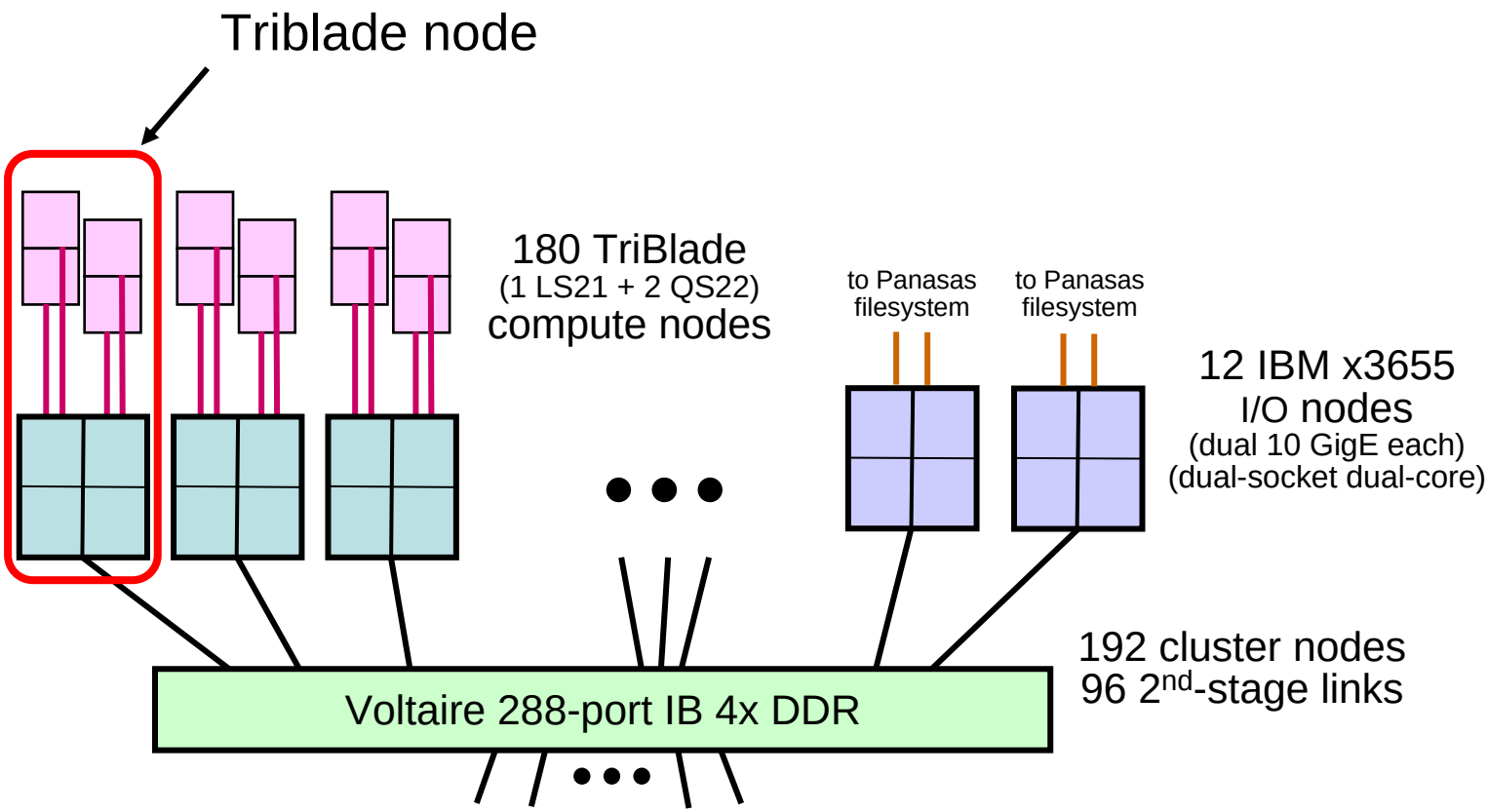
- QS22 is a future IBM Cell blade containing two new enhanced double-precision (eDP/PowerXCell™) Cell chips
- Expansion blade connects two QS22 via four PCI-e x8 links to LS21 & provides the node's ConnectX IB 4X DDR cluster attachment
- LS21 is an IBM dual-socket Opteron blade
- 4-wide IBM BladeCenter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





One Cell per Opteron core

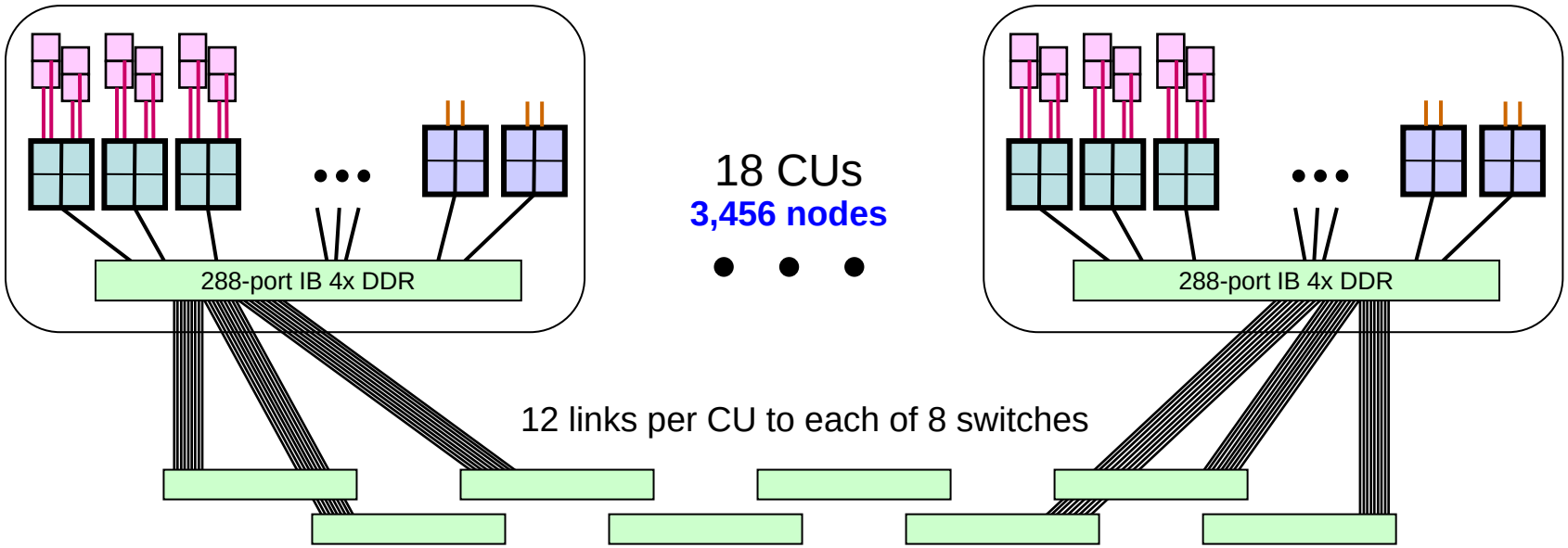
A Roadrunner Connected Unit (CU) is a semi-autonomous cluster of hybrid nodes



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

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

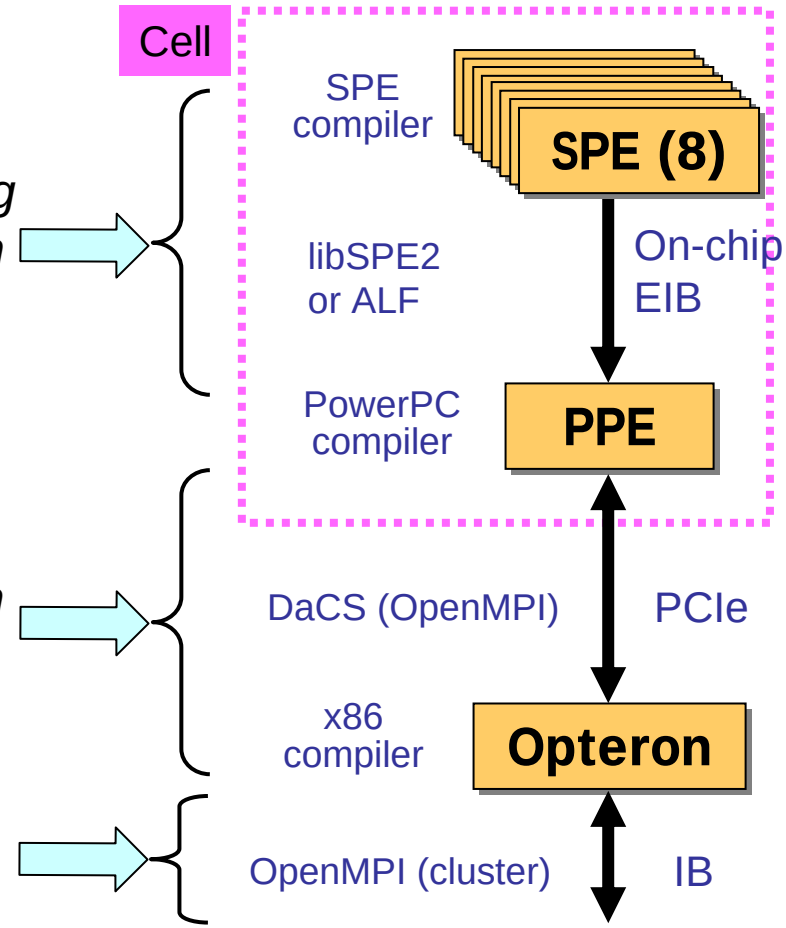
** I/O nodes not counted*



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

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



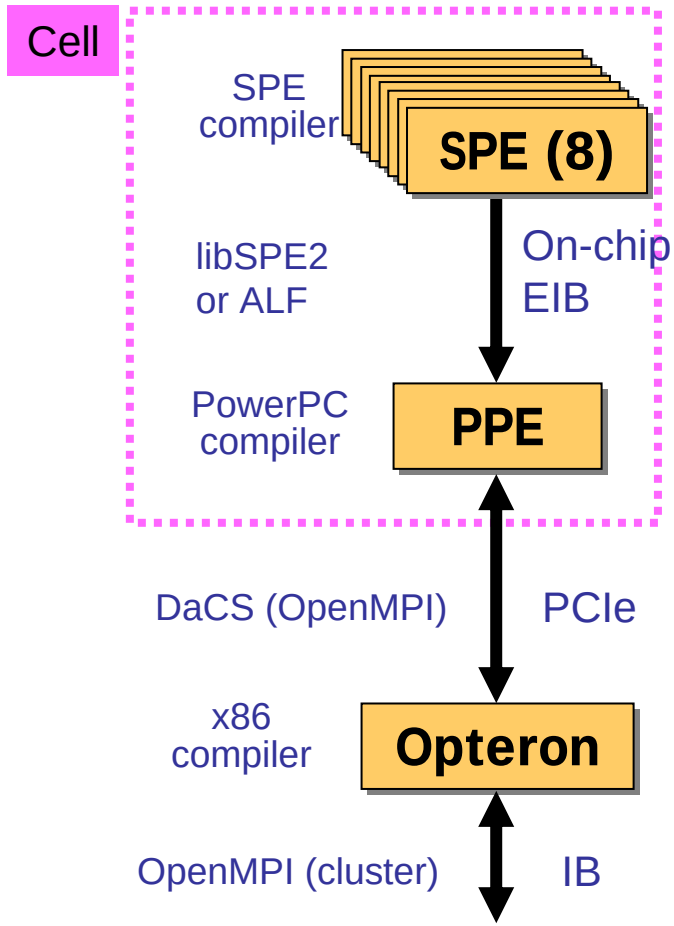
Three types of processors work together

Parallel-in-parallel
(nested parallelism)

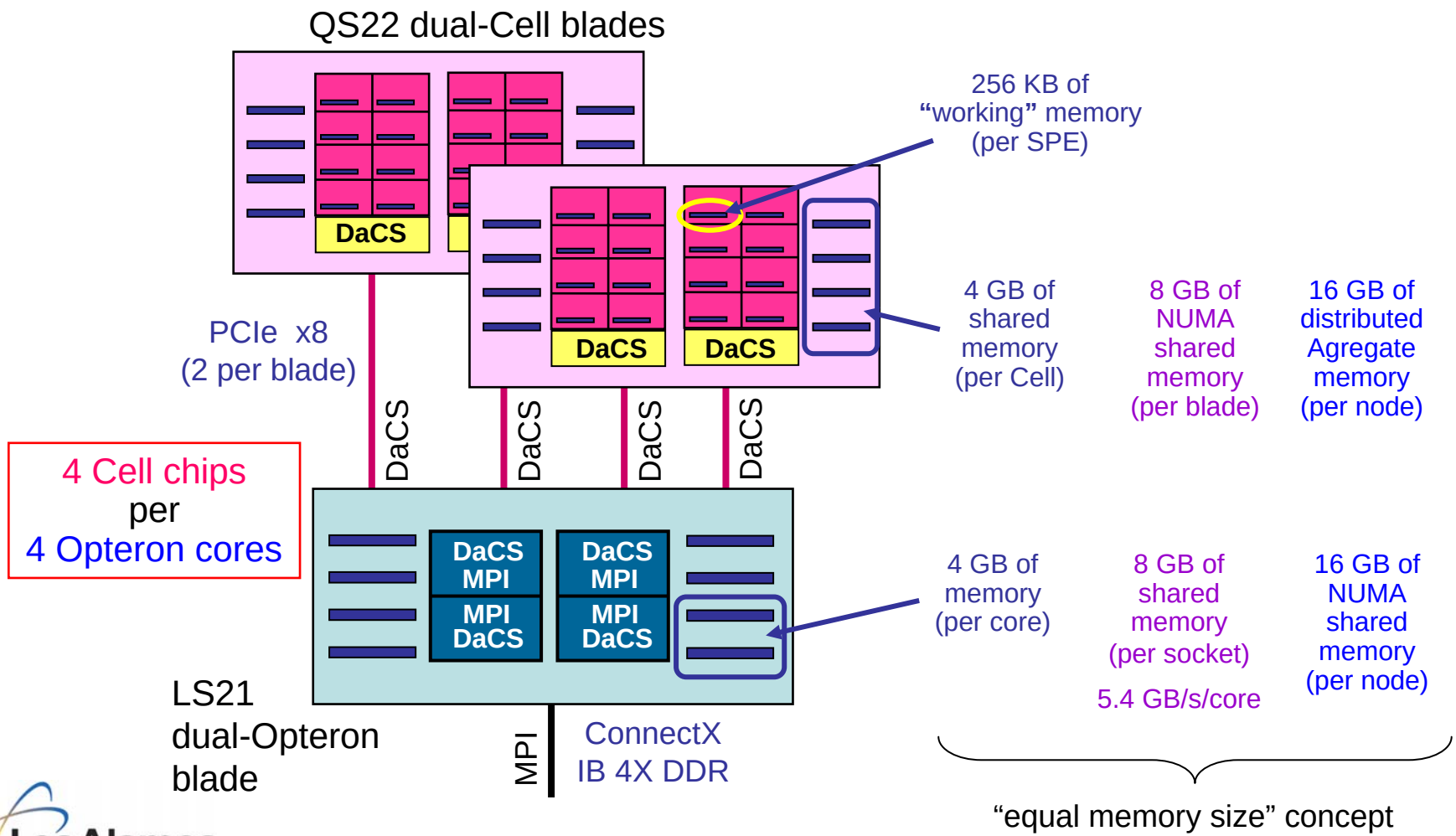
Questions to consider:

- Opteron-centric or Cell-centric control & state variables
- Data transfers (DaCS & MPI)
- Cell local store & DMAs

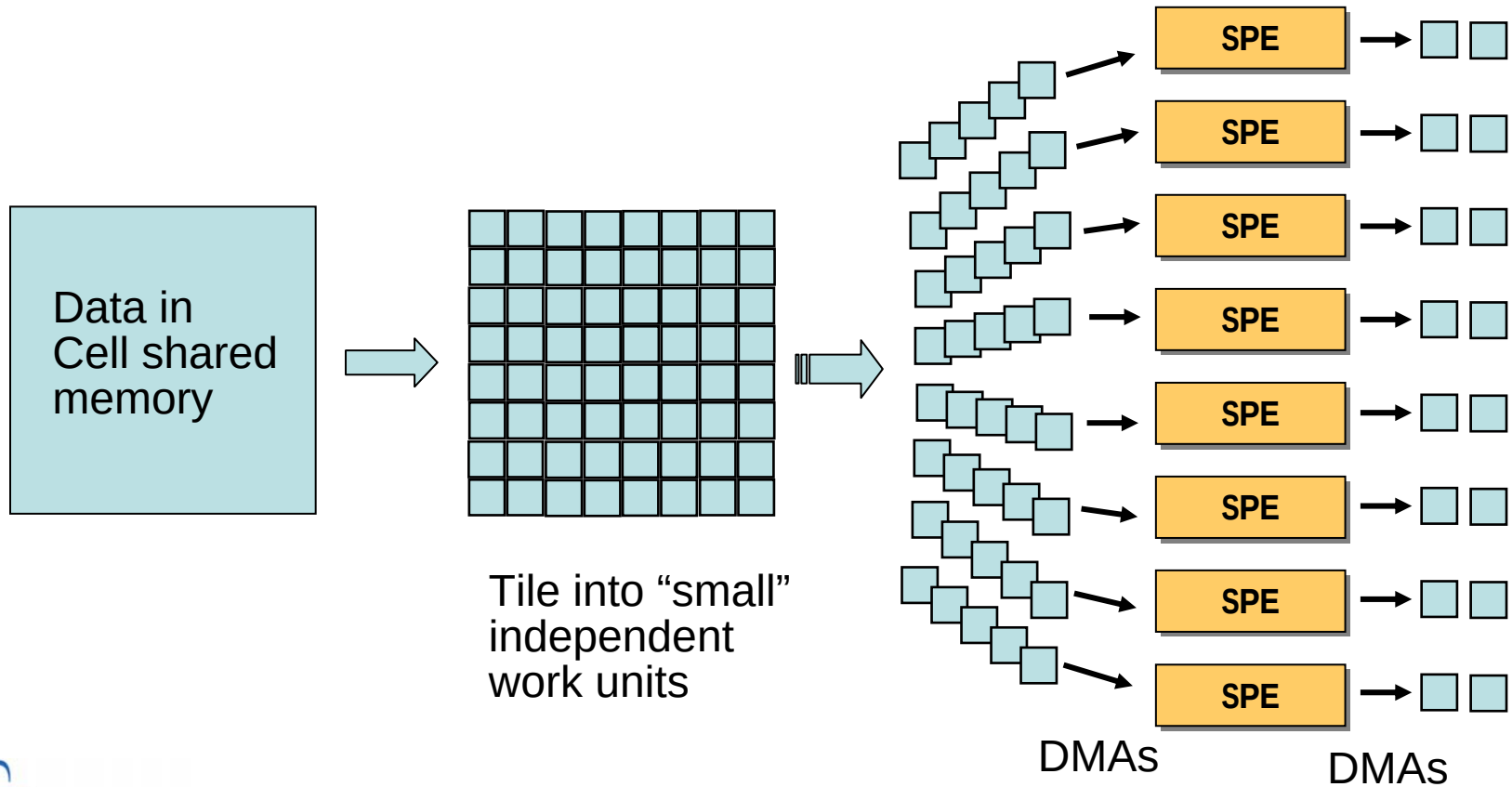
This can be done one algorithm at a time



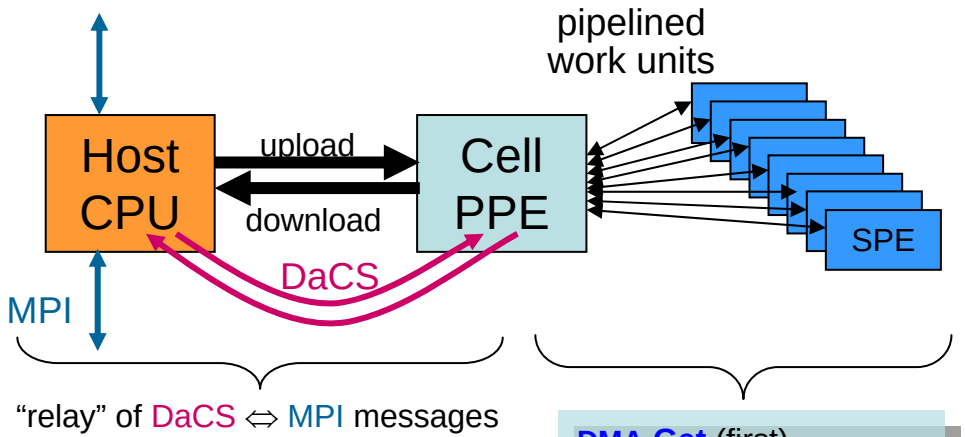
Roadrunner nodes have a memory and communications hierarchy



Cell's SPEs naturally provide shared-memory threaded parallelism



Three types of processors work 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)

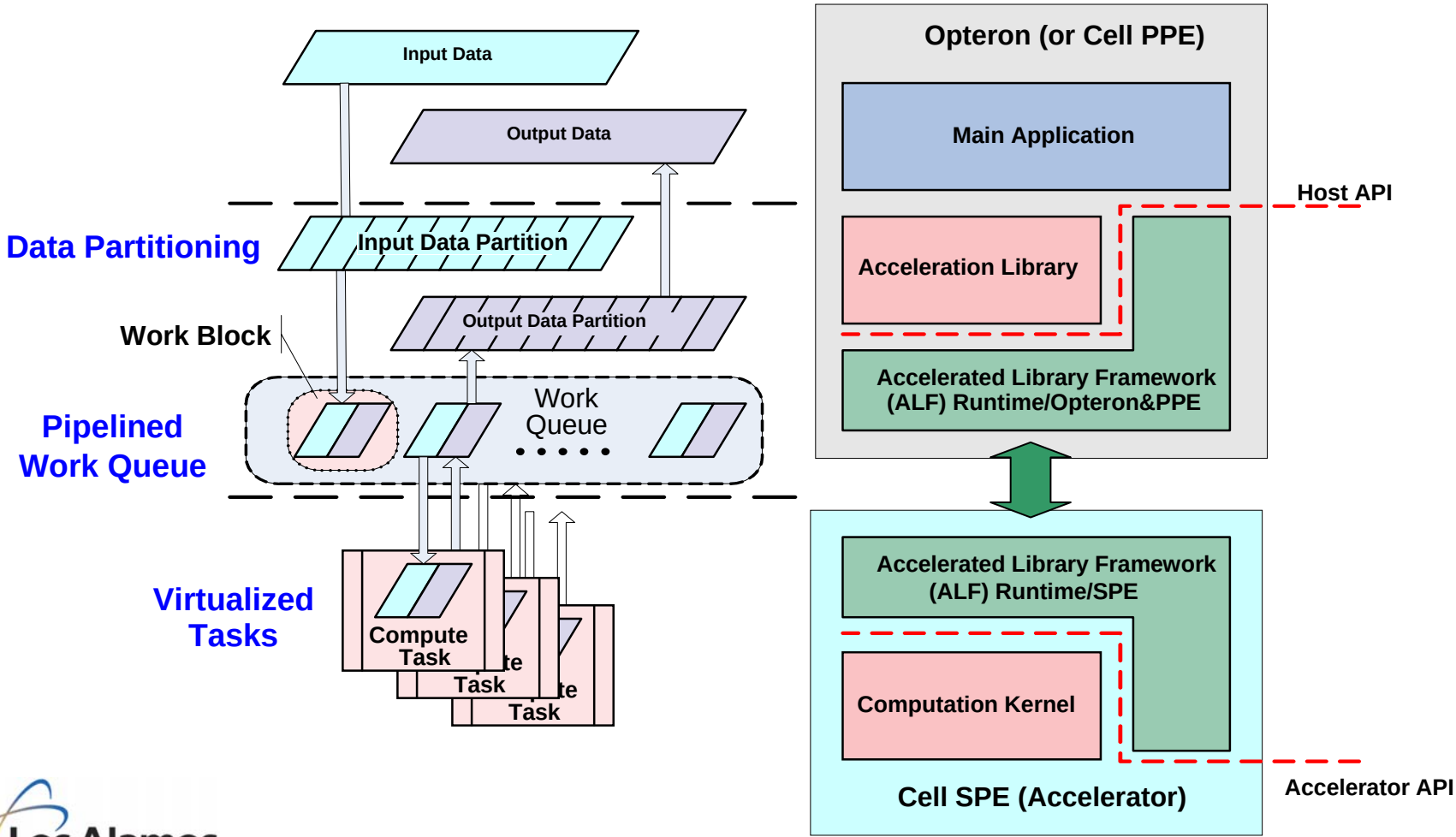
- DMAs are simply block memory transfers
  - *HW asynchronous (no SPE stalls)*
  - *DDR2 memory latency and BW performance*

DMA Get:
```
mfc_get( LS_addr, Mem_addr, size, tag, 0, 0);
```

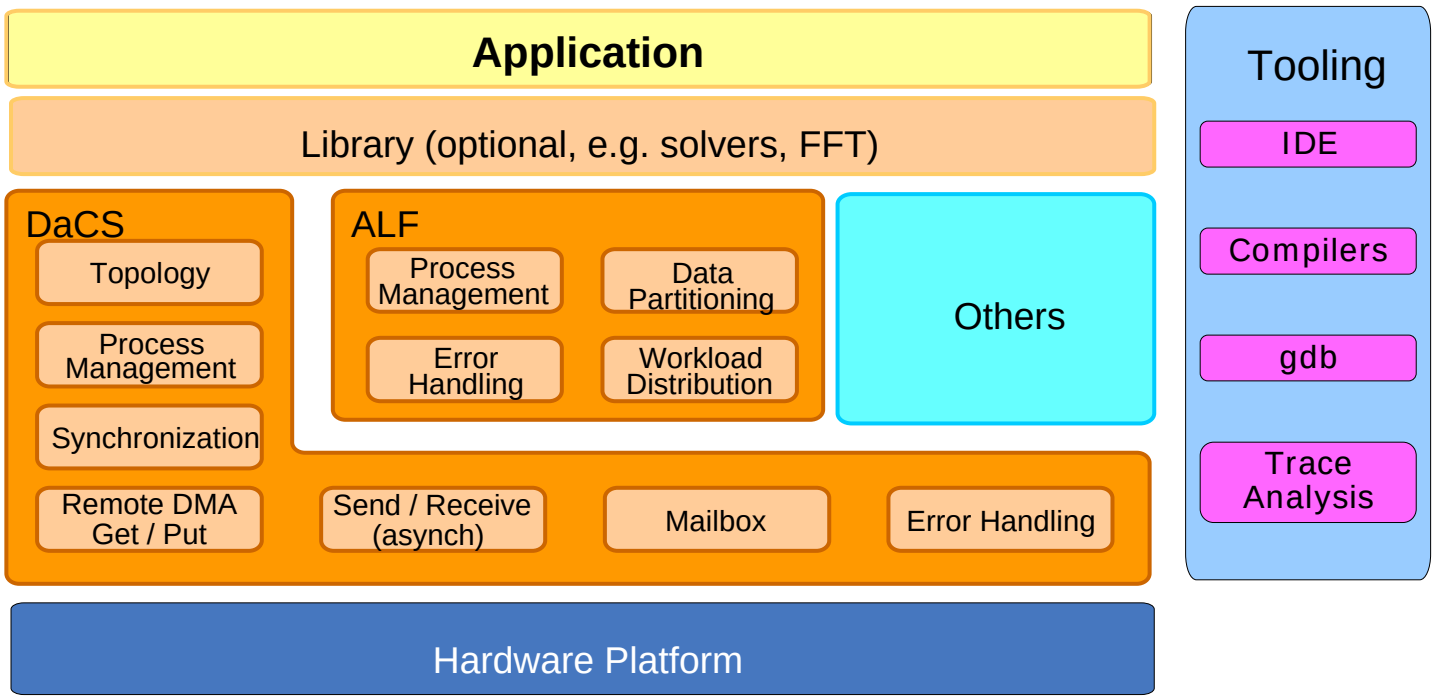
DMA Put:
```
mfc_put( Mem_addr, LS_addr, size, tag, 0, 0);
```

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

IBM-provided ALF implements a queue of work-blocks and hides scheduling & DMAs



ALF & DaCS: Broader than Cell & Roadrunner



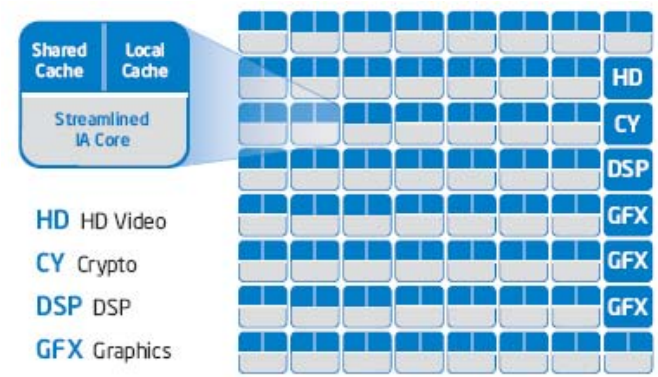
- Designed by IBM & LANL to be HW agnostic
  - multicore/GPU/Cell, interconnect, even possibly cluster-wide
  - desire technical community participation to extend range

What is “different” about programming Roadrunner? Will it continue to be different?

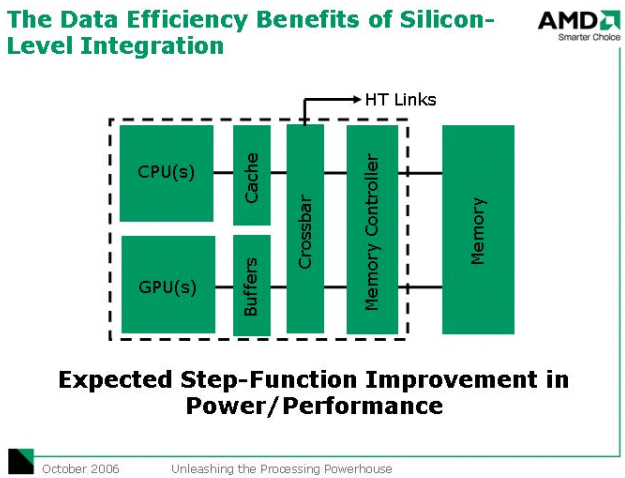
- A “rank” in an MPI parallel application becomes a set of processors with varied strengths and unequal access to memories and communication.
  - *This is similar to multi-core, then many-core and the future trends expected in micro-processors as core counts rapidly increase*
  - *Tiled work-block threaded parallelism within an MPI rank*
- Data layout and SIMD instructions are important.
  - *This is also true on all current processors for good cache and SIMD unit utilization.*
  - *Not dissimilar to previous vector codes, except now mainstream.*
- Explicit programmer management of local store.
  - *Focus effort on data locality and computational intensity, which benefit all processors*
  - *Exposes opportunities to overlap communication and computation*
  - *Non-uniform memory hierarchies is a future direction*

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

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!*

LANL Roadrunner home page

More information is available at:

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

Roadrunner Architecture

Other Roadrunner talks

Computing Trends

Related Internet links