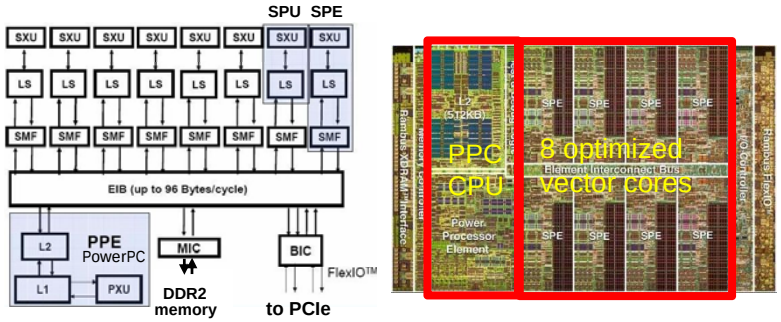
Challenges using Roadrunner

Ken Koch

Los Alamos National Laboratory

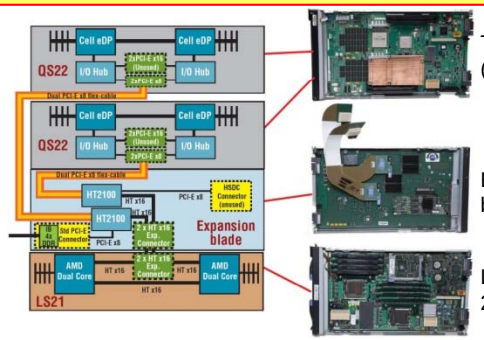
The Roadrunner Petaflop System
More at http://www.lanl.gov/roadrunner

PowerXCell 8i: an improved Cell processor



Vastly improved double precision performance
Larger DDR2-based memory

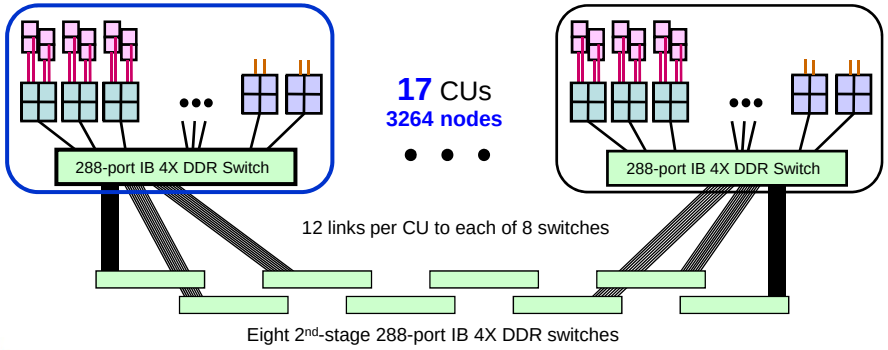
Triblade Node with PCIe-connected Cells

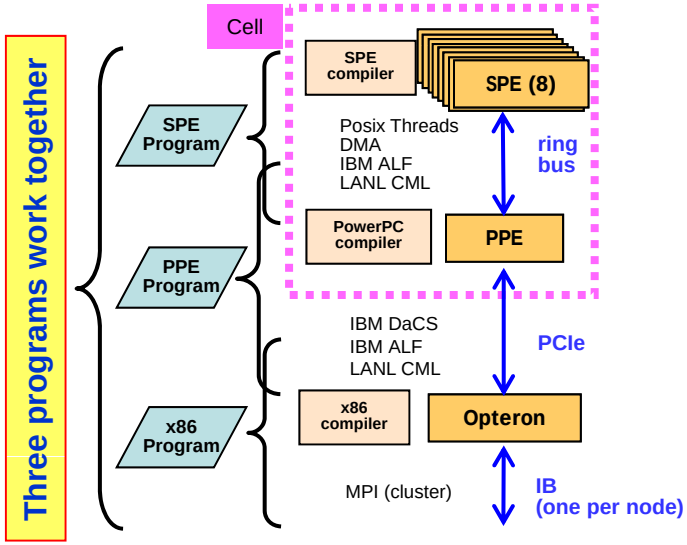


Design objective: One Cell processor for every Opteron core, plus the same memory footprint for each (4GB each), with the fastest feasible interconnects

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

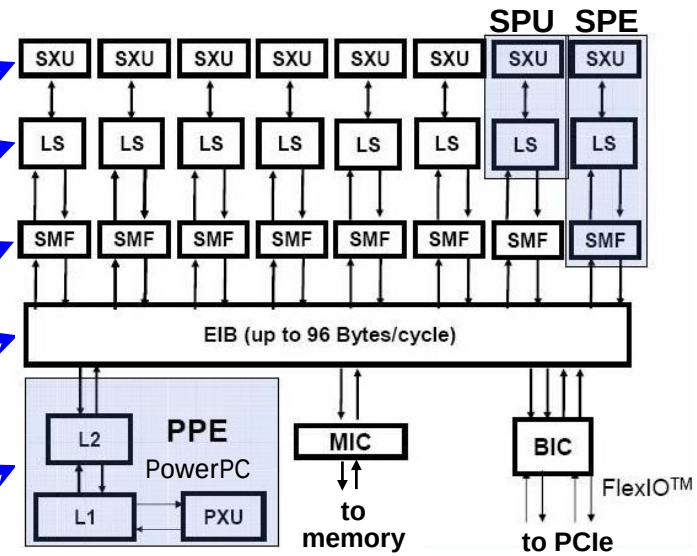
12,240 PowerXCell 8i chips ⇒ 1.33 PF, 49 TB
6,120 dual-core Opterons ⇒ 44 TF, 49 TB

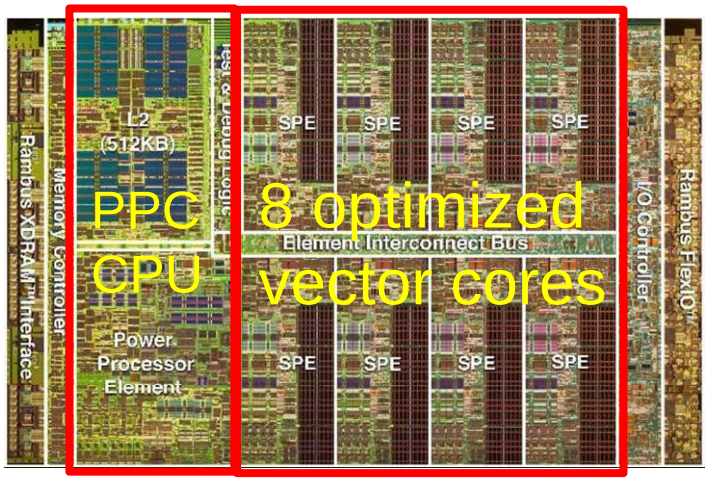




IBM created the **PowerXCell 8i**, a improved variant of the PlayStation 3 Cell processor.

- 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





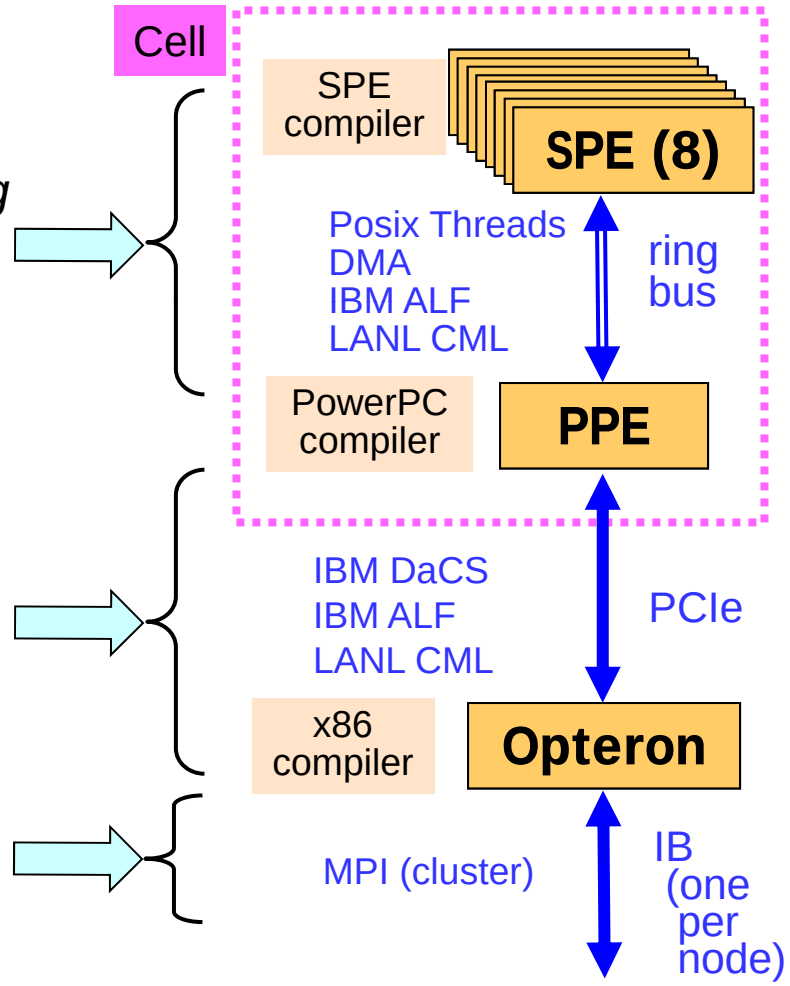
Design: SPEs provide optimal flop/s per watt in minimal area

This is an Exascale trend

* trademark of Sony Computer Entertainment, Inc.

Three types of processors are programmed to 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

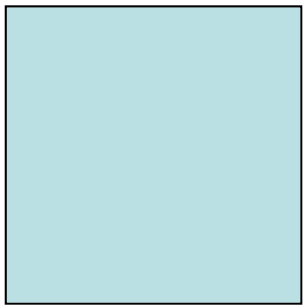


Several Challenges seen on Roadrunner

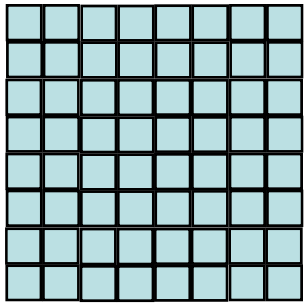
- Exposing on-node parallelism for Cell's 8 SPUs
  - *Threads are used for 8 SPUs*
- Tiling work to fit and stream in and out of the 256KB local memories
  - *Breaking up the data into chunks*
  - *Using DMA engine to overlap read-ahead/compute/write-behind*
- Breaking application into three collaborating parts (Opteron, PPC, SPUs)
  - *Most MPI programs are SPMD and bulk synchronous*
  - *Relay for MPI messages (Cell → Opteron → IB and back up)*

How do you keep the 256KB SPUs busy?

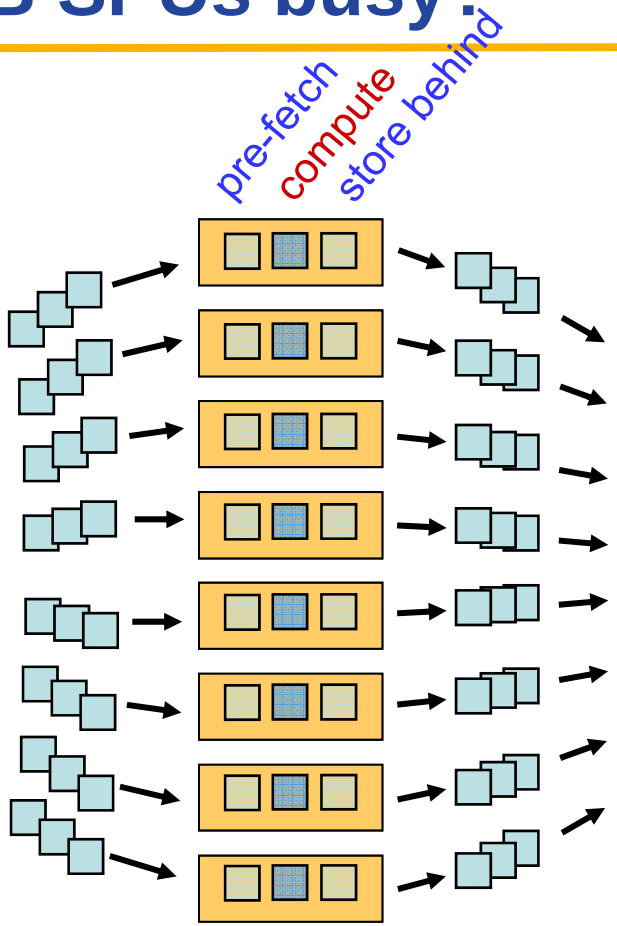
Break the work into a stream of pieces



problem domain of a Cell processor

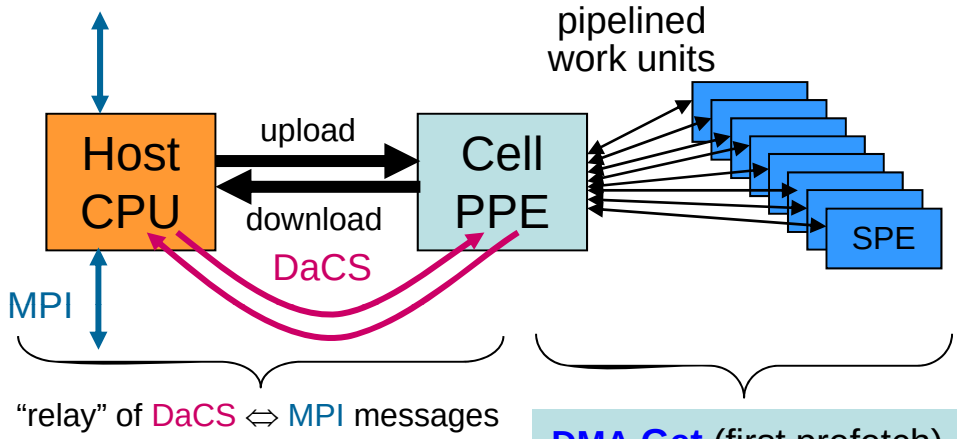


grid tiles or particle bundles (can include ghost zones)



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

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



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)

Compute & memory DMA transfers are overlapped in HW!

MPI & DaCS can also be fully asynchronous

- 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);
```

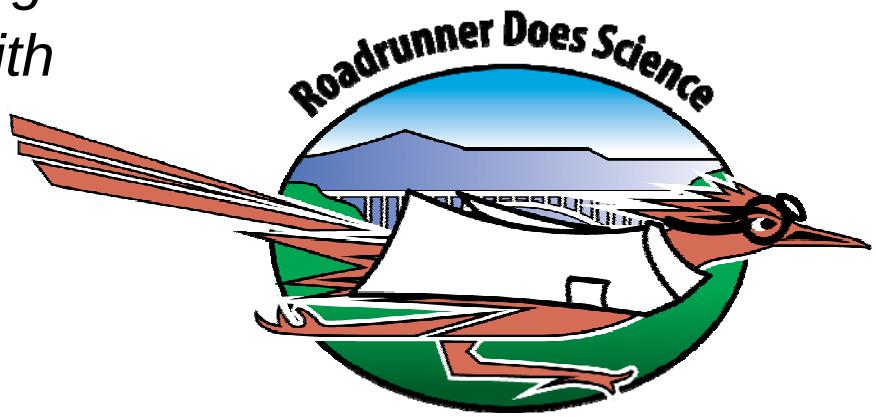
DMA Wait:
```
mfc_write_tag_mask(1<<tag);
mfc_read_tag_status_all();
```

Other Challenges

- Lack of programming tools beyond compilers
  - *Like the earliest days of parallel computing*
- Busy code developers worried about portability and longevity of an exotic platform
  - *Social acceptance*
  - *"Too busy. Too difficult. Not mainstream."*

Roadrunner Open Science

- Ten Science projects targeted Roadrunner
- February through September 2009
  - *Significant system and Panasas integration and stabilization work was ongoing during this same time*
- 7 of 10 Projects made extensive Science runs
  - *2 had good code running but ran out of time*
  - *1 ran into technical and staffing difficulties*
  - *A couple are running again with the machine in Classified mode*



The ten Roadrunner Open Science projects

Science (code)	Description	Status
Laser Plasma Instabilities (VPIC)	Study the nonlinear physics of laser backscatter energy transfer and plasma instabilities related to the National Ignition Facility (NIF).	Completed
Magnetic Reconnection (VPIC)	Study the continuous breaking and rearrangement of magnetic field lines in plasmas relevant to both space and laboratory applications.	Completed
Thermonuclear Burn Kinetics (VPIC)	Study how the TN burn process impacts the velocity distributions of the reacting particle populations and the impact that has on sustaining the burn. (ASC effort)	Code complete Open science incomplete
Spall and Ejecta (SPaSM)	Study how materials break up internally, Spall, and how pieces fly off, Ejecta, as shock waves force the material to break apart at the atomic scale. (ASC Weapons Science effort)	Mostly completed
HIV Phylogenetics (ML)	Determine "best" evolutionary relationship trees from a large set of actual genetic HIV genetic sequences (phylogenetic tree) for HIV vaccine targeting.	Completed
Properties of Metallic Nanowires (ParRep)	Apply the parallel-replica approach at the atomistic scale for simulating material properties of nanowires crucial for switches in future nanodevices.	Completed
DNS of Reacting Turbulence (CFDNS)	Study thermonuclear burning in turbulent conditions in Type Ia supernovae using Direct Numerical Simulations (DNS) with full rad-hydro.	Completed
The Roadrunner Universe (RRU)	Create a repository of particle simulations of the distribution of matter in the universe to look at galaxy-scale concentrations and structures (dark matter halos).	Mostly completed
Supernovae Light-Curves (Cassio)	Study the impact of 2D asymmetries on the radiative light output in core collapse supernovae. Coupled RAGE on Opteron-only with Jayenne-Milagro IMC (accelerated).	Code complete Open science incomplete
Cellulosomes (Gromacs)	Study the effectiveness of the decomposition of cellulosic sheets of plant fiber by cellusome bacteria related to biofuels (cellulosic alcohol) production	Code work stopped due to performance issues & manpower