Roadrunner Lessons Learned

Sriram Swaminarayan

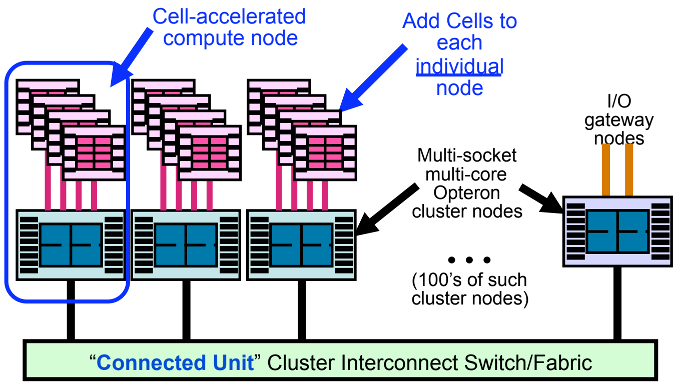
Team Leader, Evolving Applications & Architectures Team
Applications & Libraries Co-chair, Hybrid Multicore Consortium

Roadrunner at a glance.

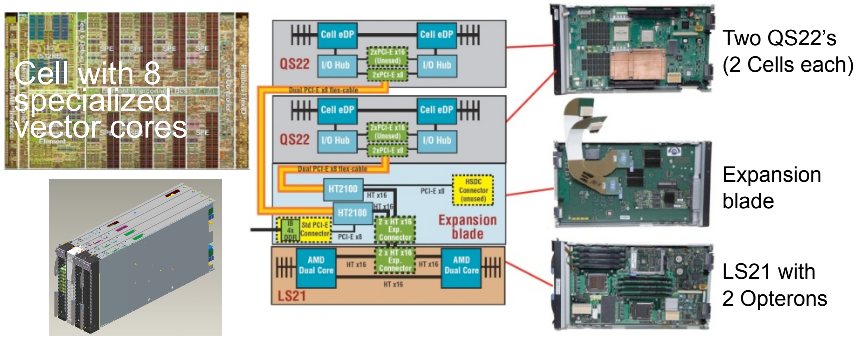
- **Cluster of 17 Connected Units (CU)**
  - *12,240* **IBM** *PowerXCell 8i chips*
  - *1.33 Petaflop/s DP peak (Cell)*
  - *1.026 PF sustained Linpack (DP)*
  - *6,120 (+408)* **AMD** *dual-core Opterons*
  - *44.1 (+4.4) Teraflop/s peak (Opteron)*
- **InfiniBand 4x DDR fabric**
  - *3264 total nodes; 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 10GigE 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 GigE network*
- **2.35 MW Power:**
  - *0.437 GF/Watt*
- **Area:**
  - *280 racks*
  - *5200 ft^2*

Roadrunner is the original (& only) hybrid petaflop supercomputer

Accelerated Node Concept


Cell-accelerated compute node
Add Cells to each individual node
I/O gateway nodes
Multi-socket multi-core Opteron cluster nodes
• • •
(100's of such cluster nodes)
"Connected Unit" Cluster Interconnect Switch/Fabric

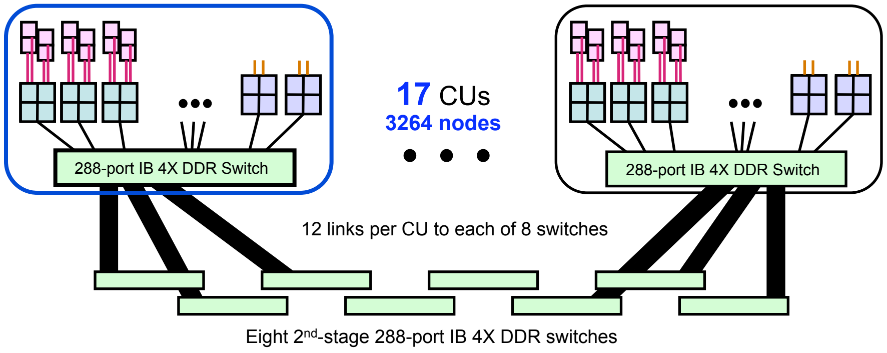
Triblade Node with PCIe-connected Cells

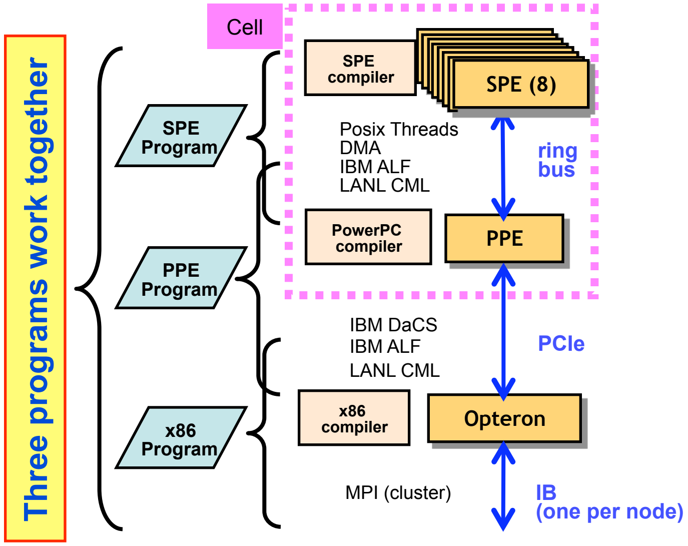


Design objective: One Cell processor for every Opteron core, plus the same memory footprint for each (4GB each), with the fastest feasible interconnects

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

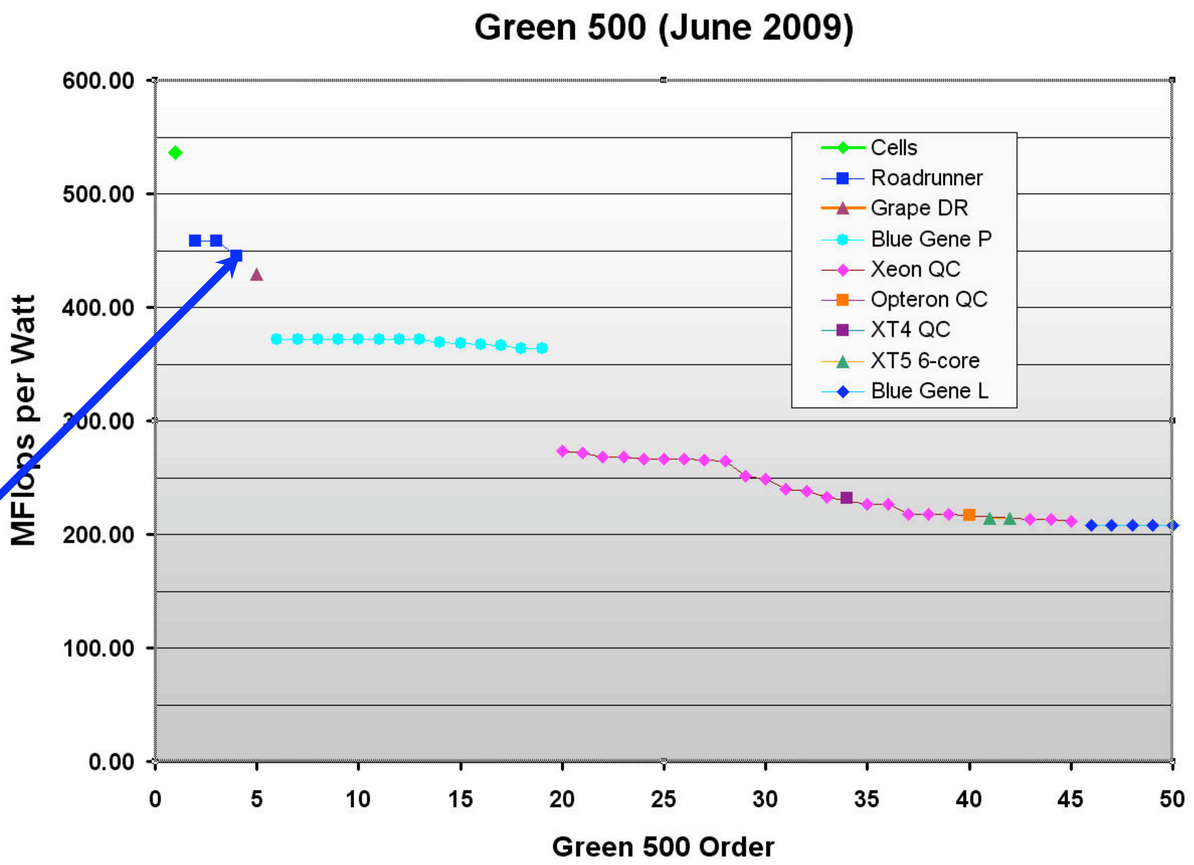

288-port IB 4X DDR Switch
•••
17 CUs
3264 nodes
• • •
288-port IB 4X DDR Switch
12 links per CU to each of 8 switches
Eight 2nd-stage 288-port IB 4X DDR switches



Roadrunner is a **Green** performer!

Now #6 (Nov 2009)

- behind 3 QPACE-PowerXCell 8i QCD clusters & the 2 single CU RRs



#4 on the Green500 (June 2009)

#2 & #3 are Roadrunner single CUs at LANL & IBM

Hybrid Vs. Conventional - 2008

Characteristic	Roadrunner	Jaguar ('08)
Peak Perf. (Pflop/s)	1.38	1.38
LINPACK % of peak	76%	77%
CPU type	Cell + Opteron (dual core)	Opteron (quad core)
Node Count	3,060	18,772
Core Count	122,400	150,176
Power (MW), measured	2.35	6.95

Same peak flop/s *but*

– 6.0× the number of nodes

– 1.2× more cores

– 3.0× the power requirement

Hybrid Vs. Conventional - 2009

Characteristic	Roadrunner	Jaguar ('08)	Jaguar ('09)
Peak Perf. (Pflop/s)	1.38	1.38	2.33
LINPACK % of peak	76%	77%	75%
CPU type	Cell + Opteron (dual core)	Opteron (quad core)	Opteron (hex core)
Node Count	3,060	18,772	18680
Core Count	122,400	150,176	224,162
Power (MW), measured	2.35	6.95	6.95

1.67× peak flop/s *but*

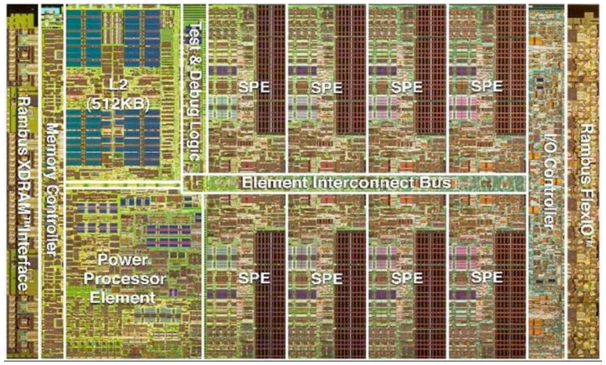
- 6.1× the number of nodes
- 1.8× more cores
- 3.0× the power requirement

Hybrid Vs. Hybrid - 2011?

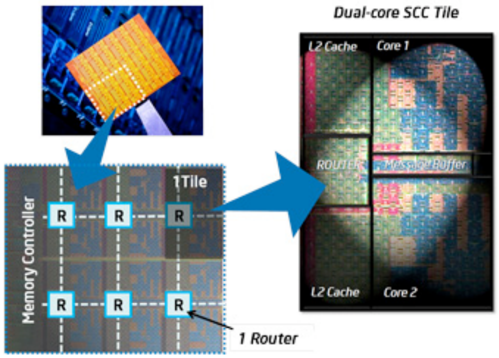
- Jaguar-11: A GPU accelerated cluster?
  - *~ 20 Petaflop accelerated cluster*
  - *Topology similar to Roadrunner!*
  - *'Traditional' Opeteron cluster*
  - *Accelerated by nVIDIA Fermi GPUs*
  - *GPUs connected over PCIe*

Roadrunner Embodies All Critical Elements Of Emerging Exascale Technologies

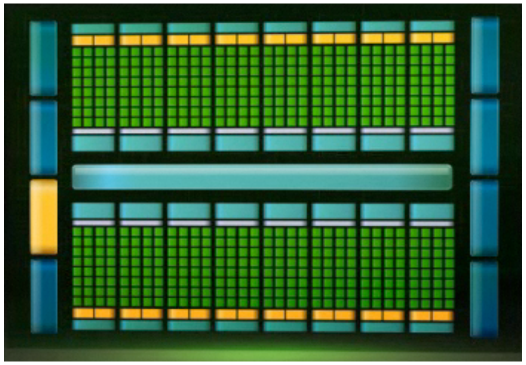
IBM Cell



Intel SCC



nVIDIA



	Opterons in Roadrunner	Opterons today (RR + 3 years)	Cell	Exascale Architectures	
				Intel	nVIDIA
Cores	2	6	8	~ 10^2	~ 10^3
Threads	4	12	8	~ 10^3	~ 10^4
Memory per core	10^7	10^7	10^5	~ 10^4	~ 10^3
Performance (GF)	10	50	100	~ 1000	~ 1000
GF/Watt	0.12	0.25	0.72	~ 50	~ 50

Taken from publicly available information

Roadrunner embodies most trends of hardware likely to be seen in the foreseeable future

- Processor Technology Issues
  - *Many Core (>16) will be the norm (e.g. Intel 48 core SCC)*
  - *Smaller, user controllable caches becoming the norm*
  - *Accelerators (e.g. GPUs from nVIDIA)*
  - *Heterogeneous chips with multiple different types of cores*
  - *Short Vector Units are the norm*
    - Need to program them directly (no magic compiler)
    - Scalar computation relegated to ‘compatibility mode’
- MPI Rank Per Core Not Sustainable
  - *Plug-in accelerators e.g. Roadrunner*
  - *MPI+threads e.g. Sequoia*
  - *Mainstream tools for on-node parallelism not MPI-aware (OpenCL, OpenMP and Grand Central Dispatch)*
- Applications will need to become more fault tolerant
  - *Large number of cores / threads implies reduced Mean Time To Failure*
  - *Strategy for codes to take hardware / software failure in their stride*
  - *Requires new strategy for checkpoint / restart*

The 10 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 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).	Partially 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

Significant speedups were obtained for hybrid applications

Project & Code	Overall Hybrid Speedup*
Laser Plasma Interaction – VPIC	~4x
Magnetic Reconnection – VPIC	2x to 3x
Spall & Ejecta – SPaSM	~5x for EAM potential
DNS of reacting turbulence – CFDNS	16x to 20x
HIV phylogenetics – ML	3x to 4.5x
RR Universe – MC^3	3x to 5x
Nanowires - ParRep	9.5x
Supernovae – Cassio	?

* Overall speedups include startup and restart I/O time

Speedup = one Cell processor + one Opteron core combined performance compared to the performance of a single Opteron core alone †

† Peak flop/s ratio = 30x; Peak memory BW ratio = 9.6x

Roadrunner Open Science Lessons Learned: Advanced Architectures Are Tractable

- Wide variety of applications have been accelerated
- A graded approach to acceleration is viable
  - *Evolutionary: 2-4x improvement*
  - *Revolutionary: 6-9x improvement*
- Data Is Everything
  - *Who owns it? (Host or Accelerator?)*
  - *Where is it now?*
  - *Where is it needed next?*
  - *How much does it cost to send it from now to next?*
- Success requires computer science experts and subject matter experts working together
- Next steps: transition knowledge to other applications

Open Science Lessons Learned – Corollary: Running In non-SIMD Mode Not Sustainable

- SIMD (Single Instruction Multiple Data) is the primary mode of computation on Roadrunner

  How fast is non-SIMD code on Roadrunner?

Cells (SPU intrinsics):	25% of peak
Opterons (SSE instructions):	50% of peak

- Accelerated computation on Roadrunner is essential for performance
  - *Cells provide the bulk of compute power on Roadrunner:*

1.325 PF = 96%	Cell
0.055 PF = 4%	Opterons

    - e.g. 20% of peak on the Opterons on Roadrunner
      - ○ *0.8% of peak on the full machine*
      - ○ ***11 TF***
      - ○ *20 TF Bottom of top500 list*

All Proposed Exascale Architectures Have Similar SIMD Characteristics

Exascale Architectures Are Roadrunner-like

- Many low power cores that embody ‘Strength in numbers’ and depend on vector instructions for performance
  - *Compare to 8 SPUs working together to provide the bulk of Cell performance*
- Billion-way Parallelism: many threads per core from 10-1000
  - *Compare to 8 SPU threads per Cell*
- Low memory per core
  - *Even less than the 256 kB/SPU on CELL*
- High bandwidth to main memory using a ring or 2D fabric
  - *Compare to the EIB Bus on Cell that provides fast access to 4G of memory*
- Focus is on performance per watt
  - *Same as Cell*
- Resilience is left as an exercise for the programmer

Considerations when moving existing applications onto next-generation platforms

- Rethinking data / algorithms will improve performance even on today's Opterons/Xeons
- Consider using higher fidelity methods
- Some changes can be done incrementally
- Changes should make the applications vector friendly
- Make changes such that we don’t paint ourselves into a corner
- Fault tolerance and resiliency issues will force memory constraints
- Best performance will require significant overhaul

Path To 'Hybridization' Requires A Layered Transition For Applications

- Target today's multicore first
  - *Make applications thread friendly (helps with Sequoia)*
  - *Requires rethinking data flow*
  - *Use Thread Manager Library to spawn threads*
- Target SIMD (SSE) next
  - *Attempt using SAL in routines that are CPU hogs*
  - *Makes debugging easy since it is done at desktop scale*
- Finally target advanced hardware with SAL bindings
  - *Use SAL bindings to compile for Cell or other architectures*

Utilizing future architectures effectively requires broader community engagement

- Building momentum in external community through partnerships with other National Labs and membership in appropriate standards bodies to guide future tool development
  - *Joint planning between NNSA(ASC) and Office of Science (ASCR)*
  - *Leadership role in DOE exascale steering committee*
    - co-chairs Rick Stevens (ANL) and Andy White (LANL)
  - *Founding member of the Hybrid Multicore Consortium (LANL,ORNL,LBNL)*
    - Executive co-chair Andy White
    - Applications co-chair Sriram Swaminarayan
    - Performance co-chair Adolfy Hoisie
  - Member of Khronos Group (governing body of OpenCL)

- Leveraging Roadrunner to advance the state of applications performance on future heterogeneous computing at exascale
  - Leveraging our unique expertise in hybrid computing built over many years leading up to Roadrunner to create applications of the future
  - Focusing on easing the transition of Applications onto next-generation architecture
  - Cerrillos, a 160 Teraflop/s, 2-CU version of Roadrunner for external collaborations (#29 on Top500 list)