LA-UR-06-xxxx

Beyond a Single Cell

Cell Workshop
University of Tennessee
October 25, 2006

Ken Koch & Paul Henning
Los Alamos National Laboratory

Roadrunner Goals

- Provide a large “capacity-mode” computing resource for LANL weapons simulations
  - Purchase in FY2006 and stand up quickly
  - Robust HPC architecture with known usability for LANL codes

- Possible upgrade to petascale-class hybrid “accelerated” architecture in a year or two
  - Follow future trends toward hybrid/heterogeneous computers
    - More and varied “cores” and special function units
  - Capable of supporting future LANL weapons physics and system design workloads
  - Capable of achieving a **<u>sustained</u>** PetaFlop

Roadrunner Phases

Stage 1 Deployment

- Phase 1 **2006**
  - Multiple non-accelerated clustered systems Oct. 2006
  - Provides a large classified capacity at LANL
  - One cluster with 7 Cell-accelerated nodes for development & testing (Advanced Architecture Initial System — AAIS)
- Phase 2: Technology Refresh & Assessment **2007**
  - Improved Cell Blades & Cell software on 6 more nodes of AAIS
  - Supports pre-Phase 3 assessment
- Phase 3
  - Populate entire classified system with Cell Blades
  - Achieve a **sustained** 1 PetaFlop Linpack
  - Contract Option

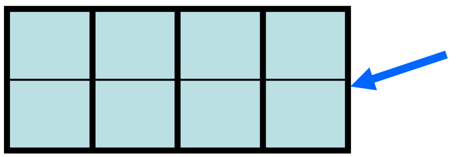
2008

Stage 2 Deployment

Base System Clusters

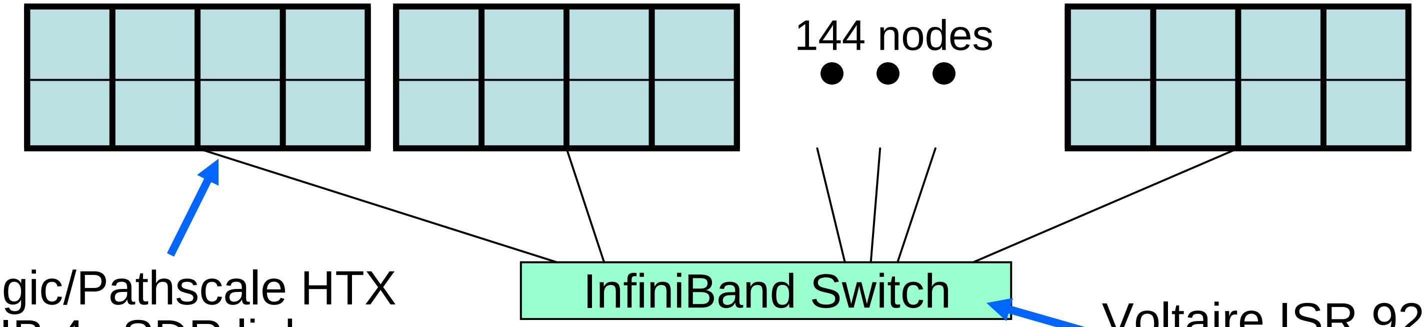
Roadrunner Connected Unit

8-way (quad-socket dual-core) Opteron Node



IBM x3755 server

Base System Connected Unit (CU) Cluster



144 nodes

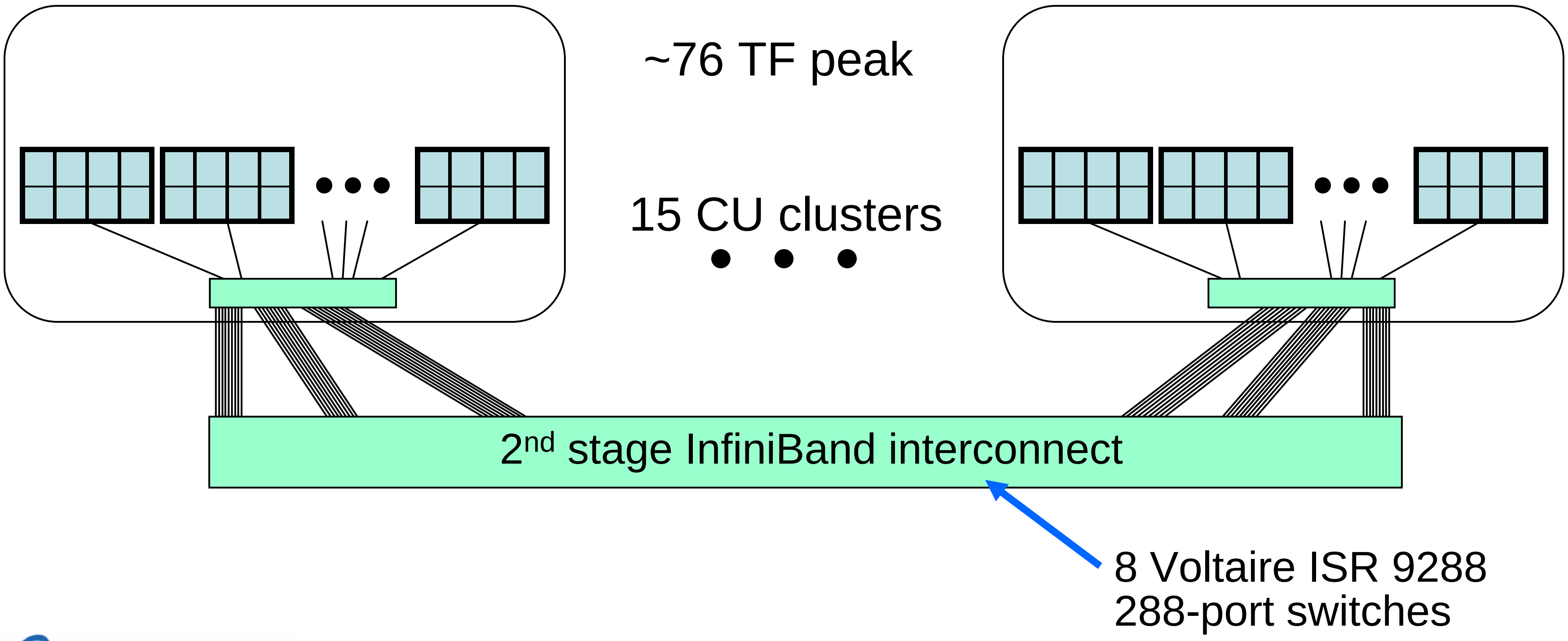
InfiniBand Switch

QLogic/Pathscale HTX
IB 4x SDR link

Voltaire ISR 9288
288-port switch

Roadrunner Base System

Multiple Cluster Base System



~76 TF peak

15 CU clusters

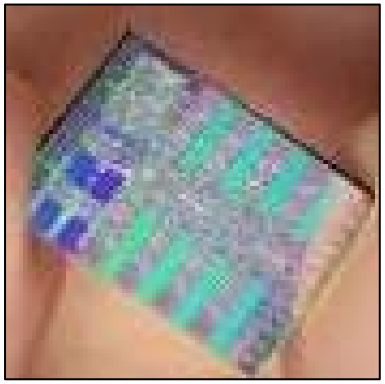
2nd stage InfiniBand interconnect

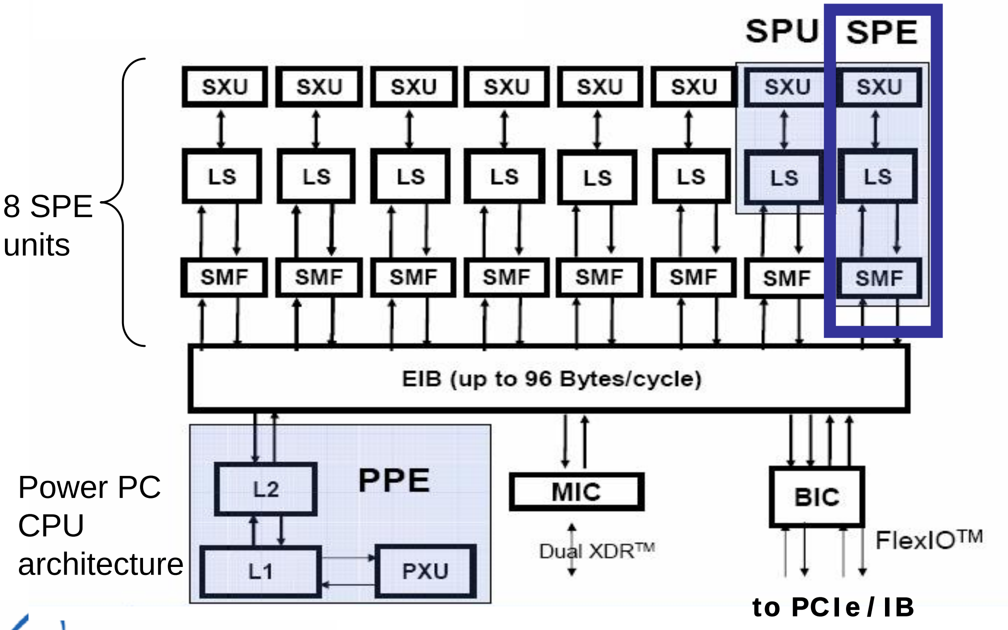
8 Voltaire ISR 9288
288-port switches

Cells as Accelerators

Cell Chip

- Cell Broadband Engine™ * (Cell BE)
  - Developed under Sony-Toshiba-IBM efforts
  - Current Cell chip is used in the Sony PlayStation 3
- An 8-way heterogeneous parallel engine





Each of the 8 SPEs are 128 byte (e.g. 2-way DP-FP) vector engines w/ 256KB of Local Store (LS) memory & a DMA engine.

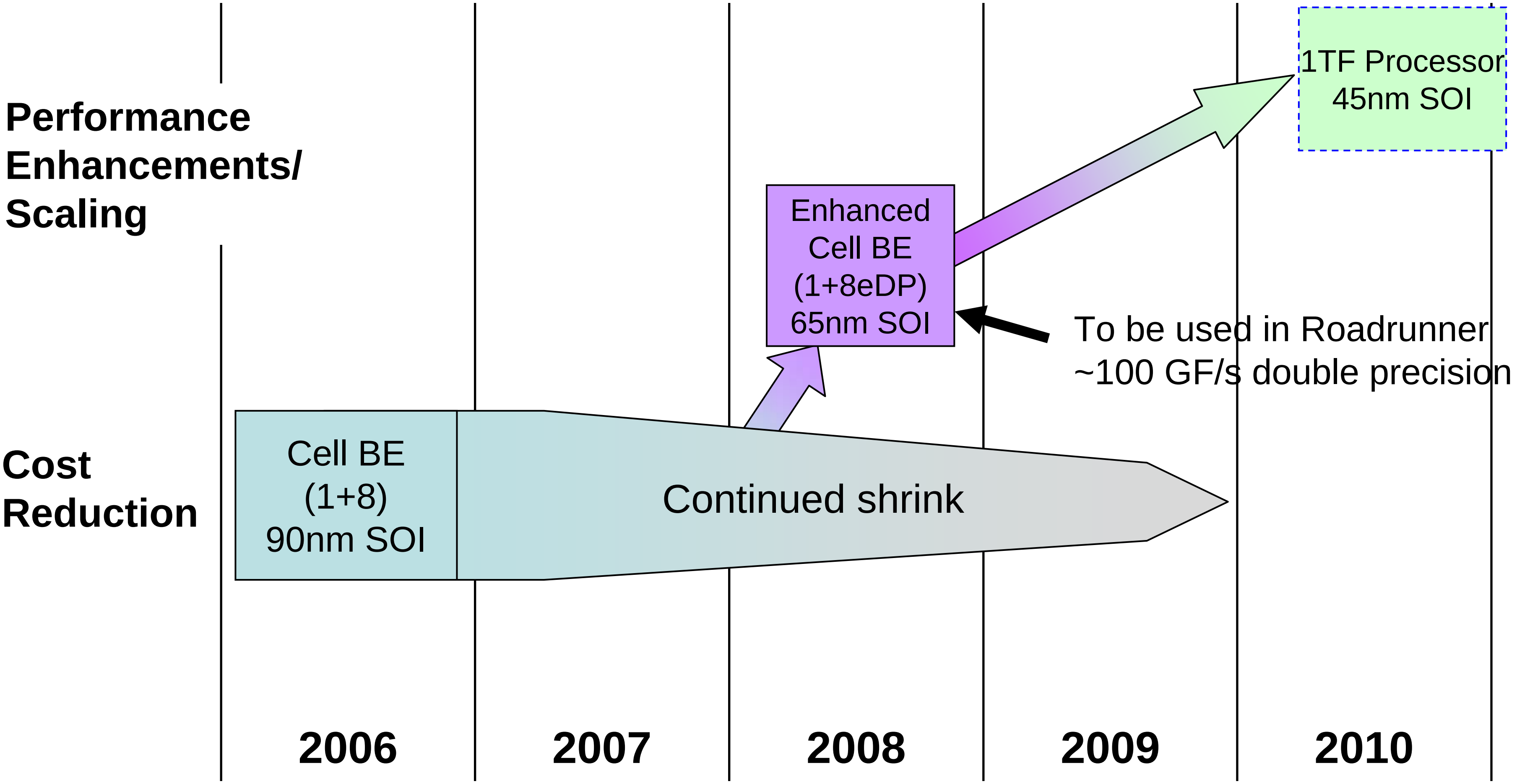
They can operate together or independently (SPMD or MPMD).

~200 GF/s single precision

~ 15 GF/s double precision (current chip)

* Trademark of Sony Computer Entertainment, Inc.

Cell Broadband Engine Architecture™ Technology Competitive Roadmap



Performance Enhancements/ Scaling

Cost Reduction

Cell BE (1+8) 90nm SOI

Continued shrink

Enhanced Cell BE (1+8eDP) 65nm SOI

1TF Processor 45nm SOI

To be used in Roadrunner ~100 GF/s double precision

2006

2007

2008

2009

2010

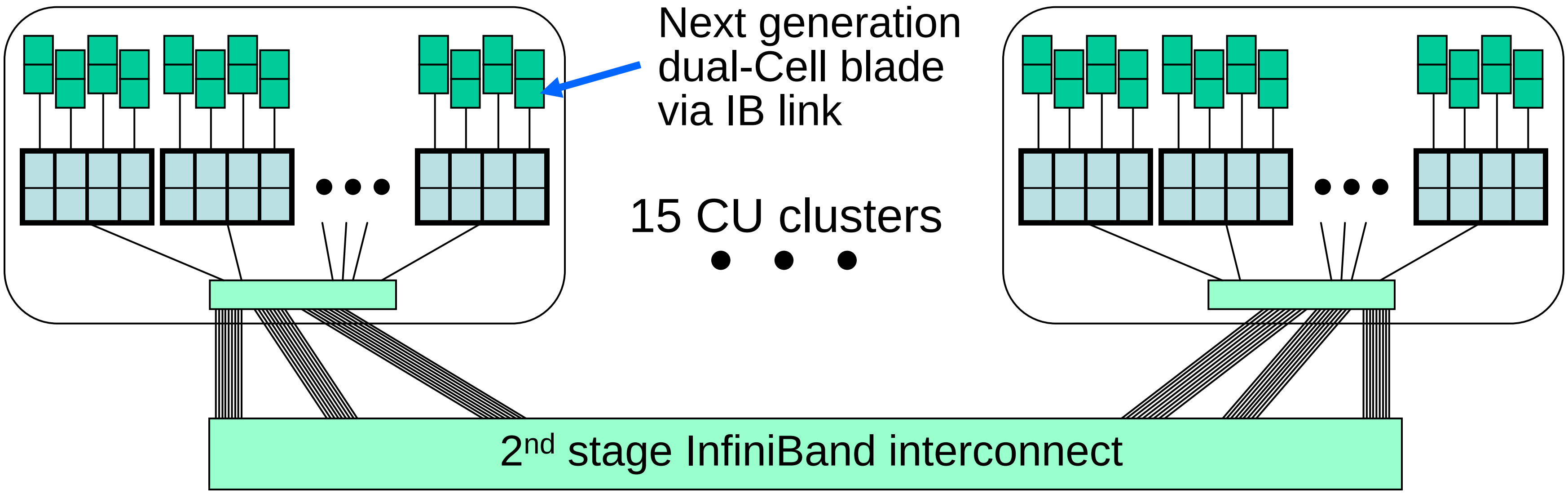
Cell BE Roadmap Version 5.0 24-Jul-2006

All future dates are estimations only; Subject to change without notice.

Roadrunner with Cells

Final System with Cell Blade Accelerators

~1.7 PF peak or Cell double precision



Next generation dual-Cell blade via IB link

15 CU clusters

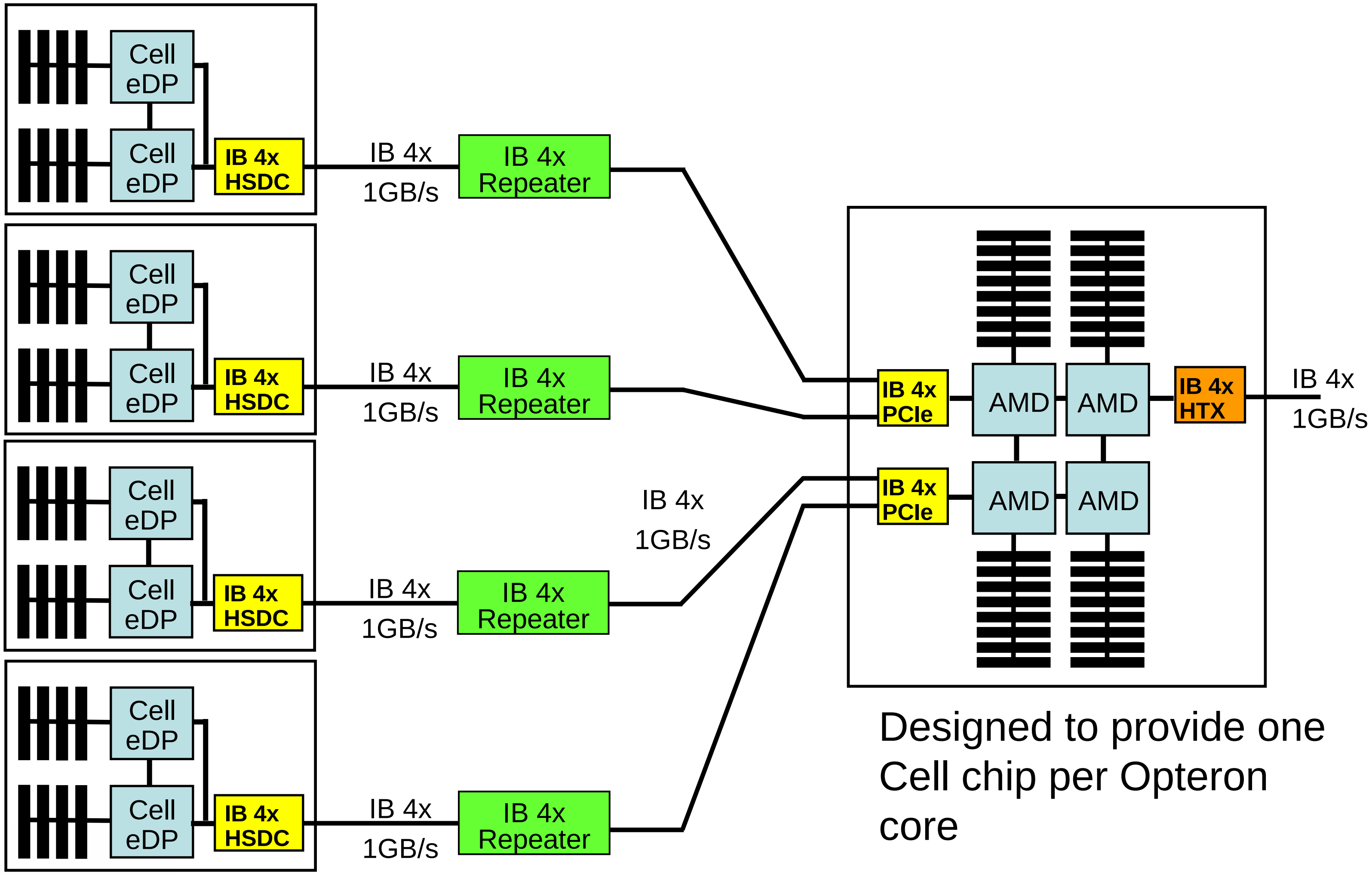
● ● ●

2nd stage InfiniBand interconnect

Cell blades are attached via direct IB links to 138 nodes of each CU

16,560 total eDP Cell chips in the Phase 3 Roadrunner accelerated system

Accelerated Node

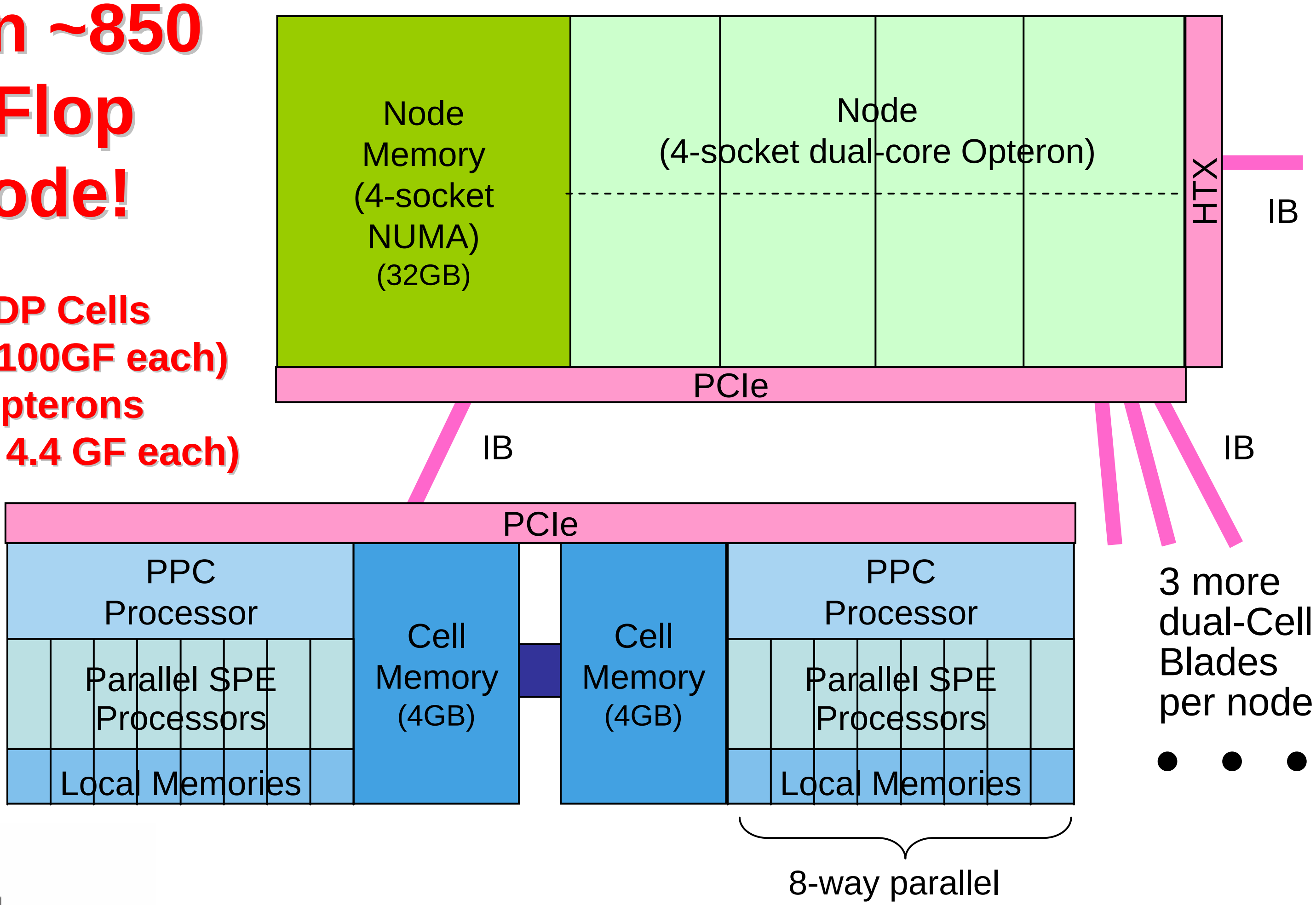


Designed to provide one Cell chip per Opteron core

Roadrunner Heterogeneity

An ~850 GFlop Node!

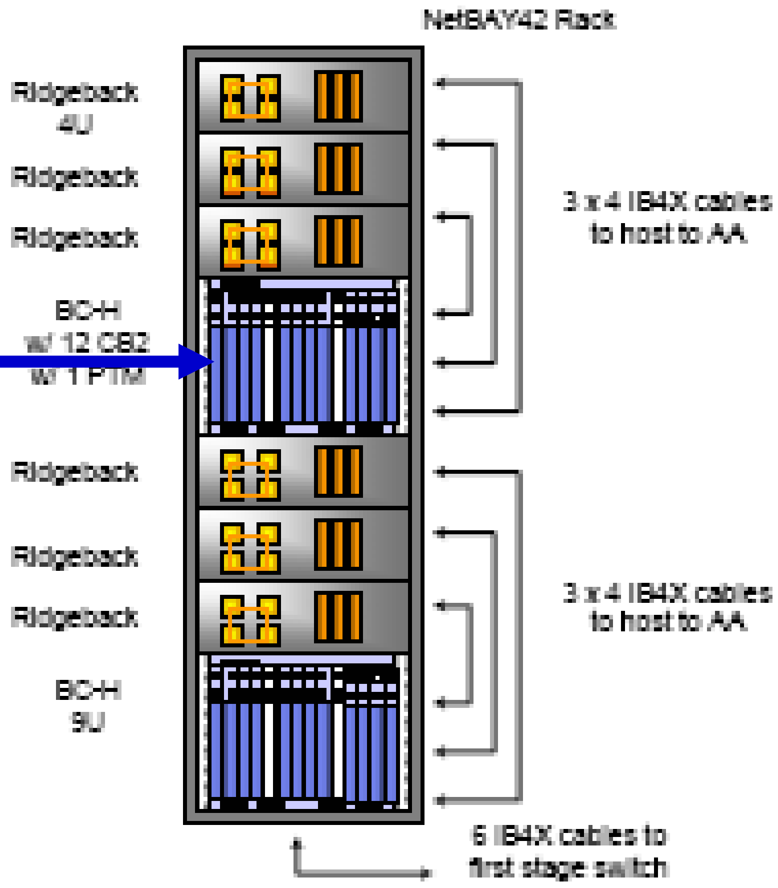
8 eDP Cells (~100GF each)
8 Opterons (~ 4.4 GF each)



Compute Rack

12 blades

BladeCenter-H

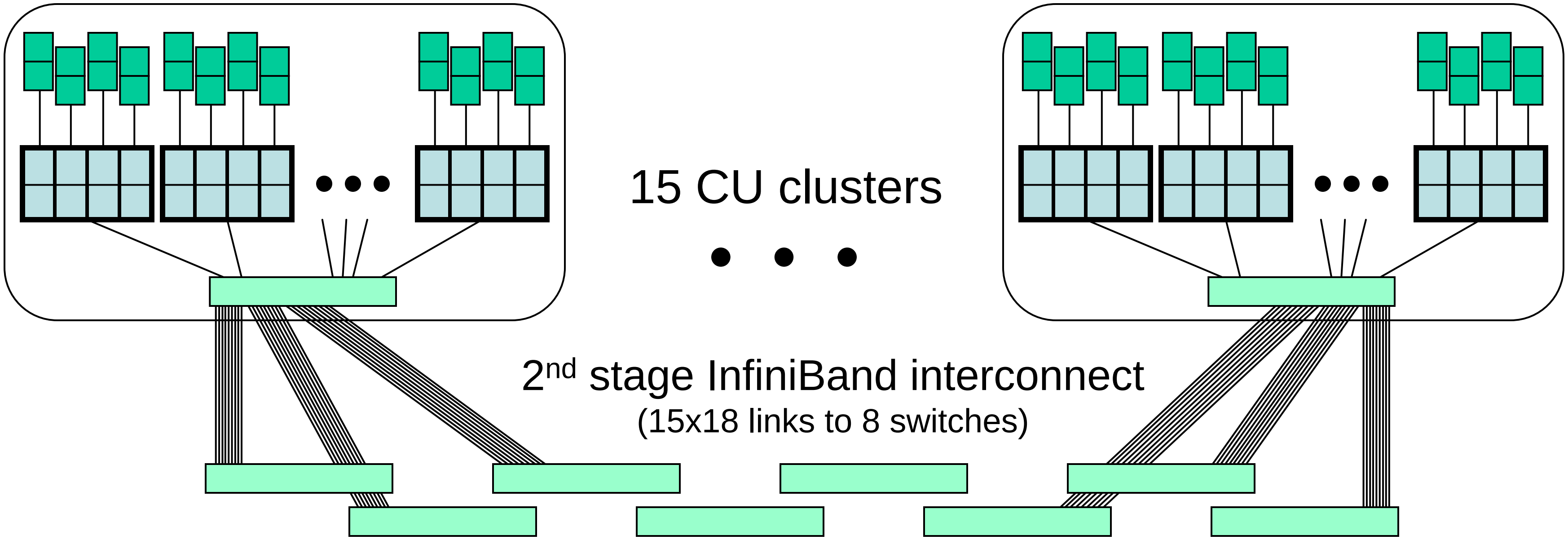


~16 KW per rack
~1 KW per x3755 Ridgeback
~5 KW per BC-H w/ 12 Cell Blades

Accelerated Roadrunner

“Connected Unit” cluster
144 quad-socket
dual-core nodes
(138 w/ 4 dual-Cell blades)
InfiniBand interconnects

In aggregate:
8,640 dual-core Opterons + 16,560 eDP Cell chips
76 TeraFlops Opteron + ~1.7 PetaFlops Cell



15 CU clusters

2nd stage InfiniBand interconnect
(15x18 links to 8 switches)

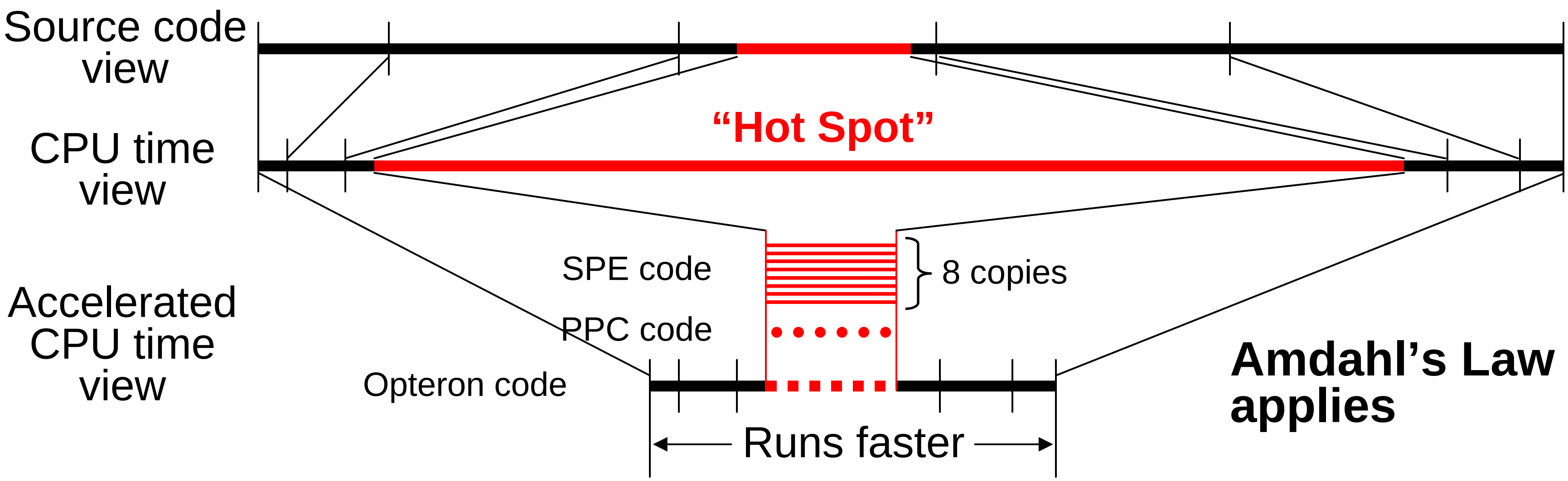
Hybrid Programming

- Roadrunner is hybrid/heterogeneous
  - Standard Opteron-only parallel codes run unaltered on Roadrunner cluster nodes
  - Computationally intense kernels or entire modules or pieces are partially modified or rewritten to take advantage of Cells
    - Hopefully limit the source code impacted
- A hybrid code would have 3 distinct cooperating pieces
  1. Main code runs on Opteron of a node
  2. A Cell PPC code
  3. A Cell SPE code
  - Developer architects the cooperation now; tools may be able to help some in the future

Hybrid Programming

- Decomposition of an application for Cell-acceleration
  - Opteron code
    - Runs non-accelerated parts of application
    - Participates in usual cluster parallel computations
    - Controls and communicates with Cell PPC code for the accelerated portions
  - Cell PPC code
    - Works with Opteron code on accelerated portions of application
    - Allocates Cell common memory
    - Communicates with Opteron code
    - Controls and works with its 8 SPEs
  - Cell SPE code
    - Runs on each SPE (SPMD) (MPMD also possible)
    - Shares Cell common memory with PPC code
    - Manages its small Local Store (LS) memory, transferring data blocks in/out as necessary
    - Performs vector computations from its LS data
- Each code is compiled separately (currently)

Cell Programming



- Hybrid programming will be a challenge!
  - No compiler switches to "just use the Cells"
  - Not even a single compiler – 3 of them
  - Code developer/architect must decompose application and create cooperative program pieces

Opteron-Cell Programming Environment

- Minimum requirements:
  - Job launch & control, including delivery of executable image
  - I/O and error forwarding
  - Asynchronous data communication, DMA & MP styles
    - Double-buffered data transfers with computation
  - Synchronization primitives
- "Simple" Leverage Approach is Open MPI, but it…
  - Doesn't deliver executables to Cell Blades
  - Currently has some lingering problems with heterogeneous MPI_Comm_spawn()
    - Opteron->PPC
  - Makes attached accelerator explicit
    - 2 levels of communications

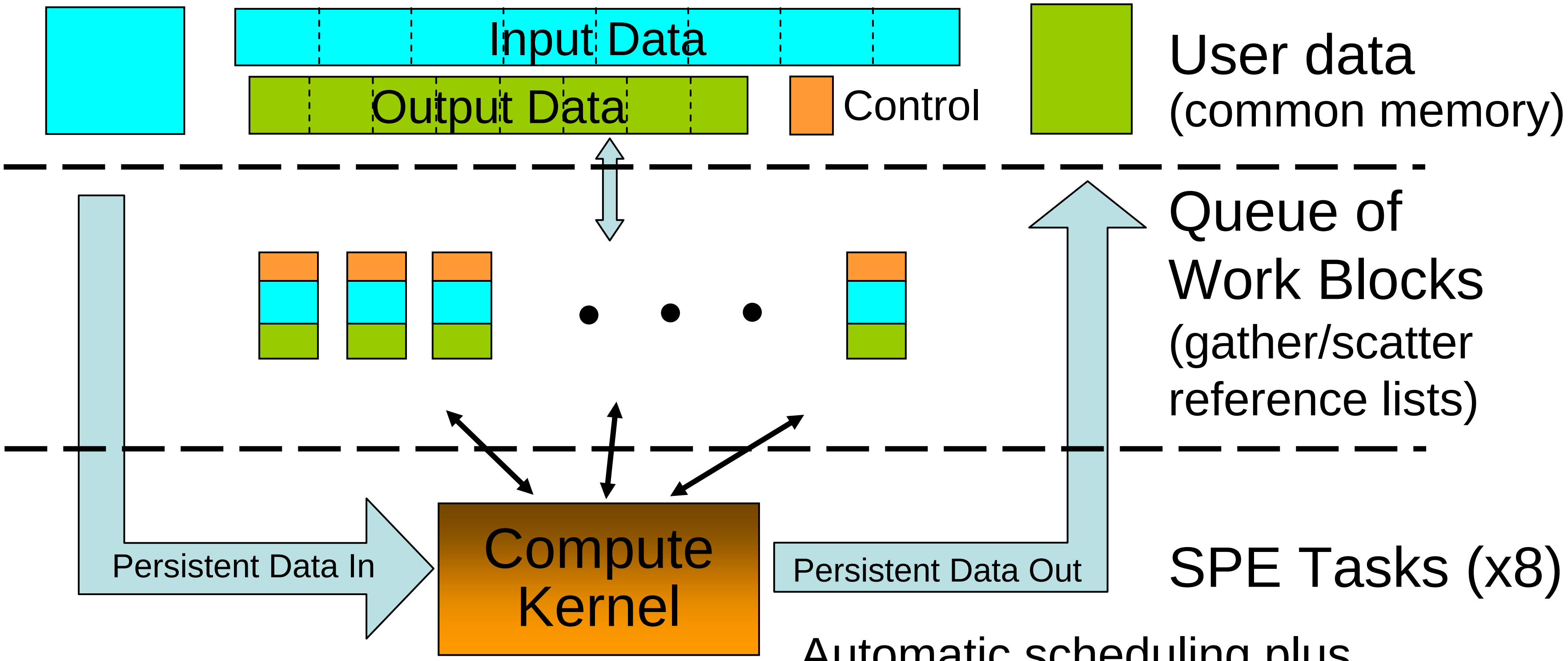
IBM/LANL Communication API

- API being developed to meet minimum requirements.
  - Support Roadrunner's IB connected Cell Blades
  - Primarily in C, but is friendly to C++ and F9x
- Hides the particulars of the interconnect fabric
  - more future-proof.
- Processor topology and reservation system
  - Allows precise process placement for MPMD
  - Good for managing communications links and NUMA issues
  - Adapts to future hardware configurations
- Not specific to Cell or Roadrunner

Work Queue API

- High-level API
  - Should be good for data-parallel operations
  - Option to programming to the hardware using low-level intrinsics
- Implements a common communication paradigm to increase programmer productivity and robustness
- Automatically partitions work among accelerators.
- Overlaps DMA operations with compute kernel
- No extra data copies
  - Working data defined by gather/scatter lists

Work Queue Paradigm



Input Data

Output Data

Control

User data
(common memory)

Queue of
Work Blocks
(gather/scatter
reference lists)

Persistent Data In

Compute
Kernel

Persistent Data Out

SPE Tasks (x8)

Automatic scheduling plus
overlap of communication &
computation

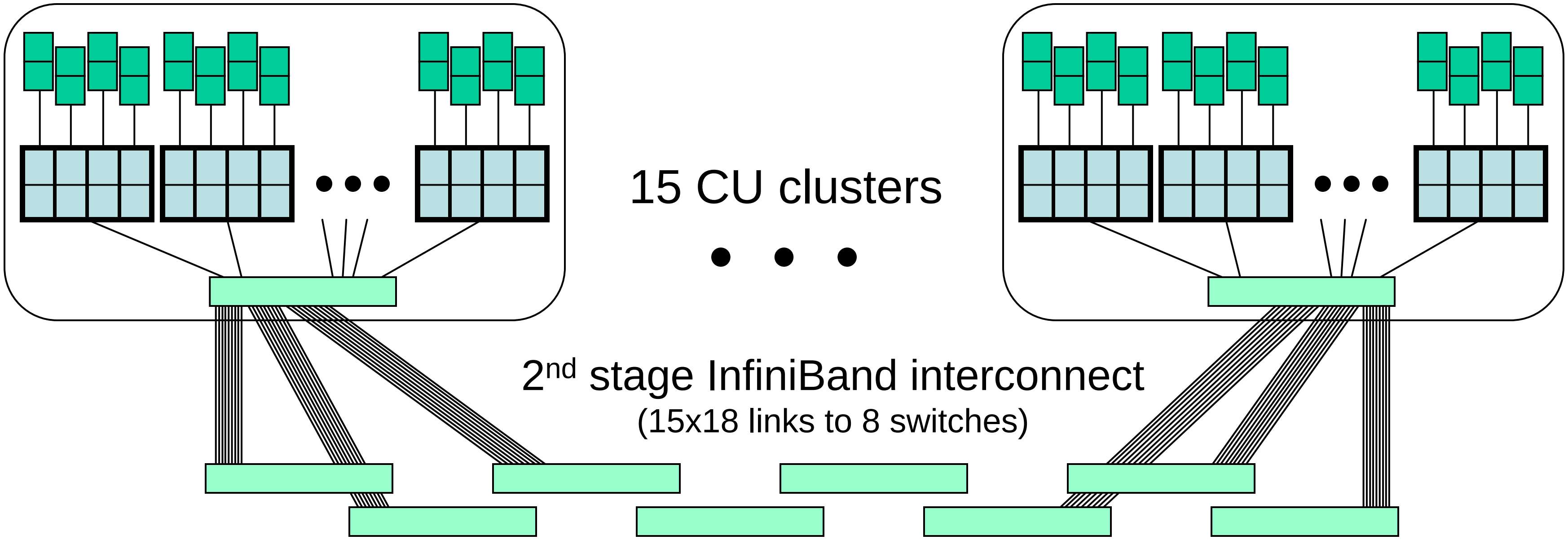
Thank you for your attention

Questions & Answers?

Accelerated Roadrunner

“Connected Unit” cluster
144 quad-socket
dual-core nodes
(138 w/ 4 dual-Cell blades)
InfiniBand interconnects

In aggregate:
8,640 dual-core Opterons + 16,560 eDP Cell chips
76 TeraFlops Opteron + ~1.7 PetaFlops Cell



15 CU clusters

2nd stage InfiniBand interconnect
(15x18 links to 8 switches)