OPENFABRICS ALLIANCE

RDMA in Embedded Fabrics

Ken Cain, kcain@mc.com
Mercury Computer Systems
06 April 2011

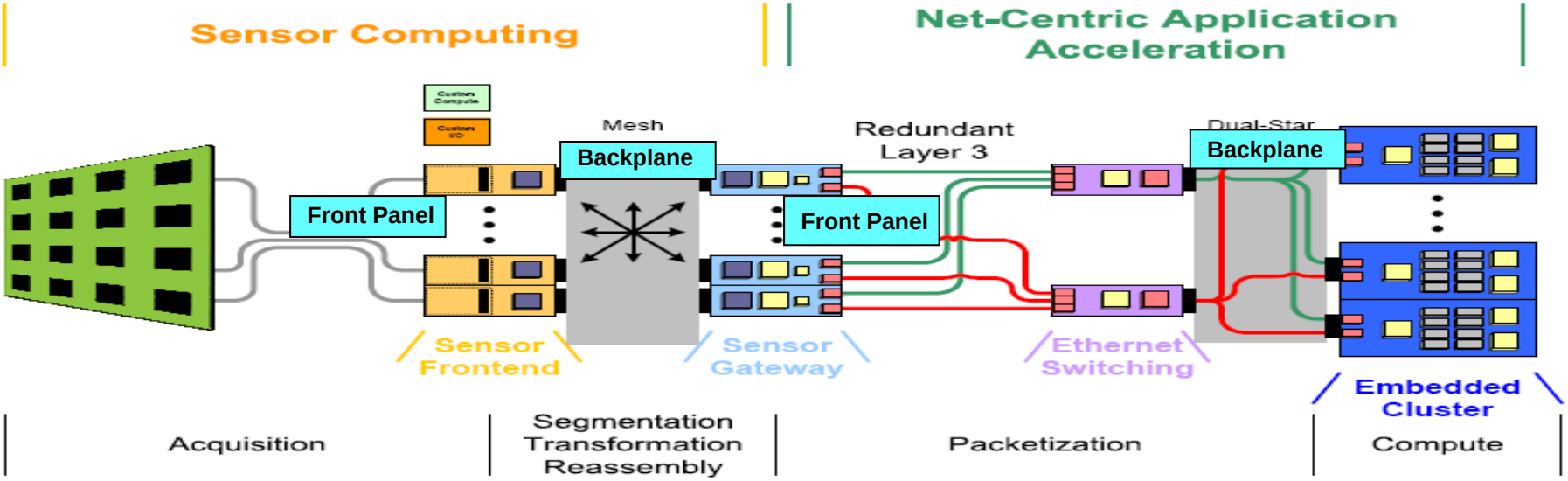
Outline

- Embedded Systems Architecture / OpenVPX
- Open MPI, OFED for Serial RapidIO (sRIO)
- Embedded Data Flow Use Cases
- OFED Enhancement Opportunities
- Conclusions

Embedded Systems Architecture Model

- OpenVPX

Typical Embedded System Architecture

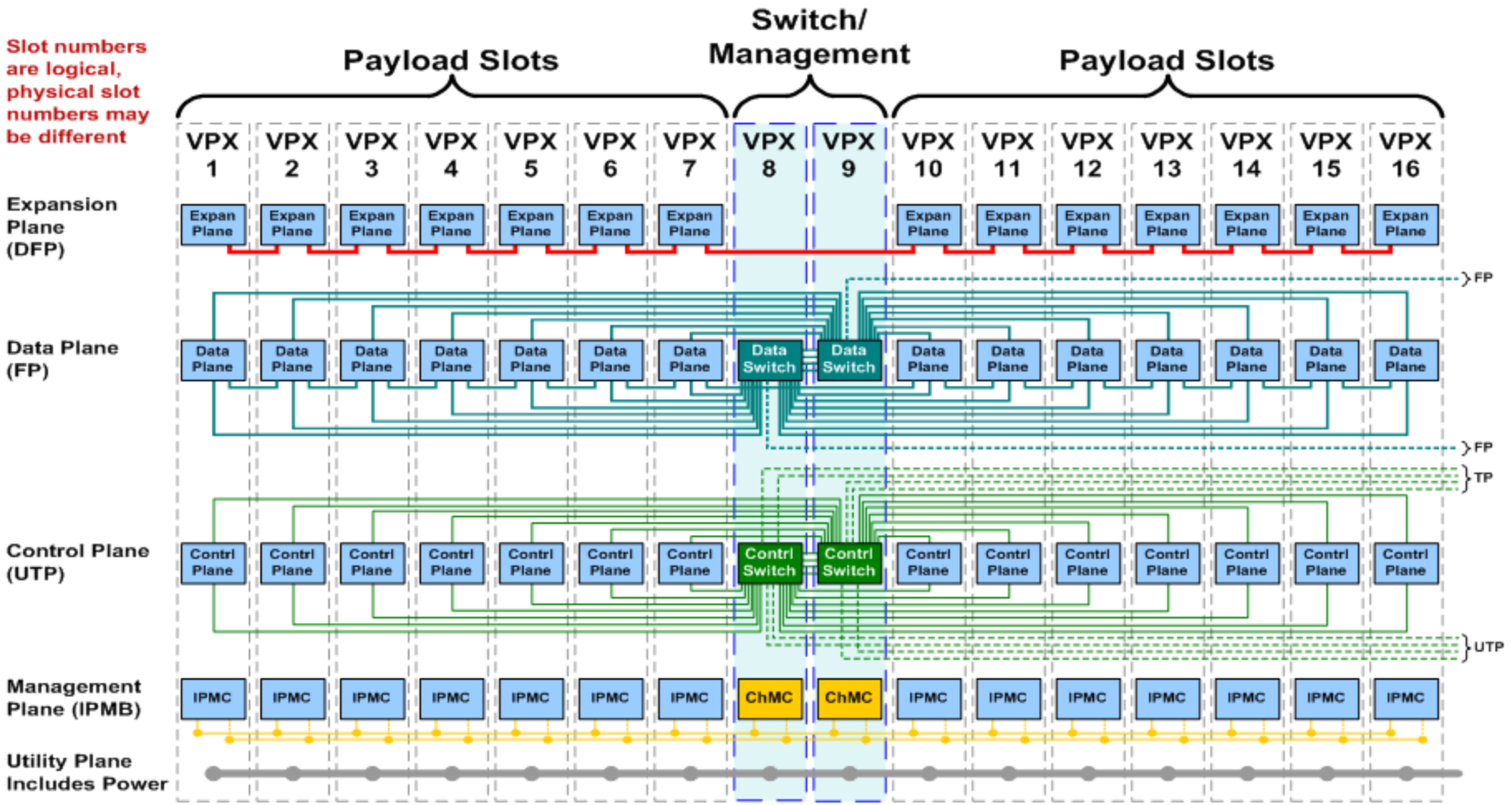


RADAR, sonar, video, etc. sensors
Sensor compute modules (FPGA)
A to D converters, RF tuners

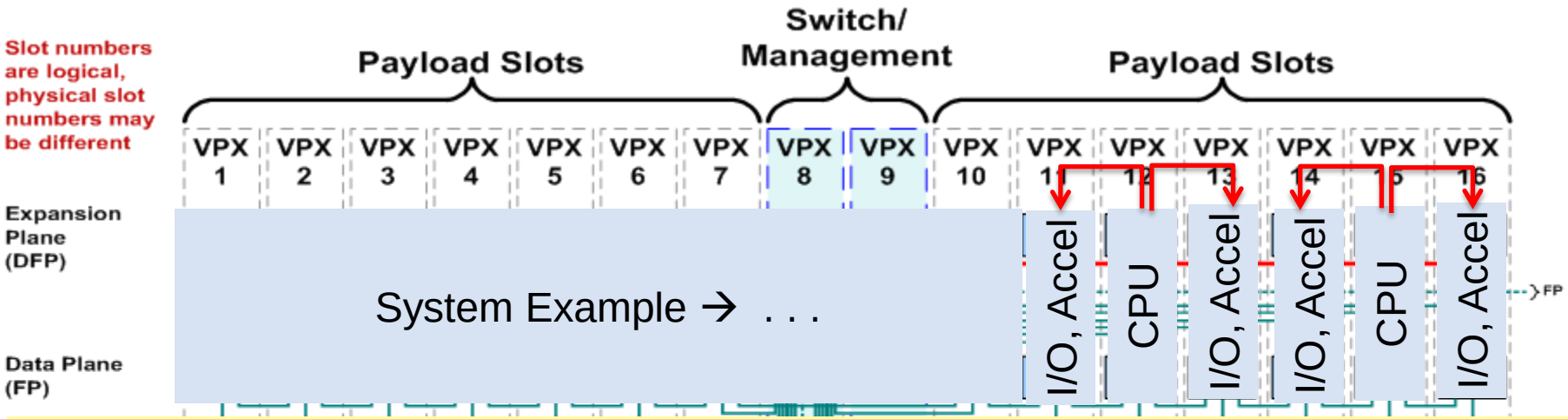
VITA 41, 46/48, OpenVPX – 3U, 6U, SBC modules
DSP modules (PPC / x86 multicore, GPU)
Switch module (e.g., serial RapidIO – sRIO)

- **Dataflows among sensor I/O and heterogeneous compute devices**
- **Size/weight/power constraints – "embedded servers?"**
- **Compute, fabric efficiency → smaller (or more capable) systems**

OpenVPX Model



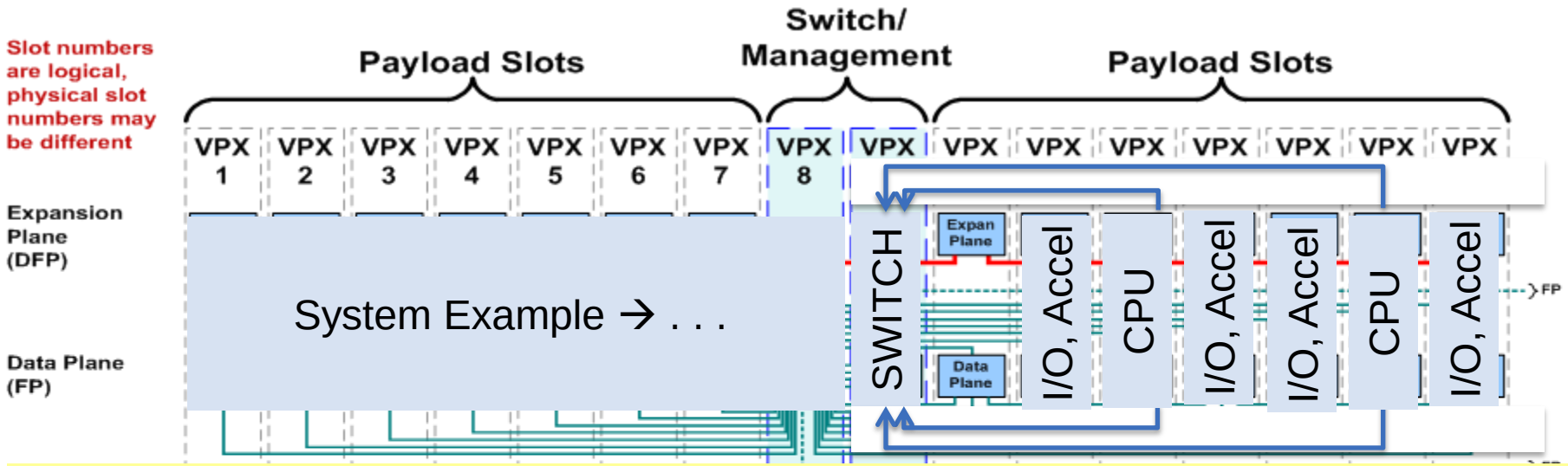
OpenVPX Model



Connectivity – Expansion Plane

- PCIe to Adjacent Slots
- Examples: GPGPU, FPGA Sensor/Compute

OpenVPX Model



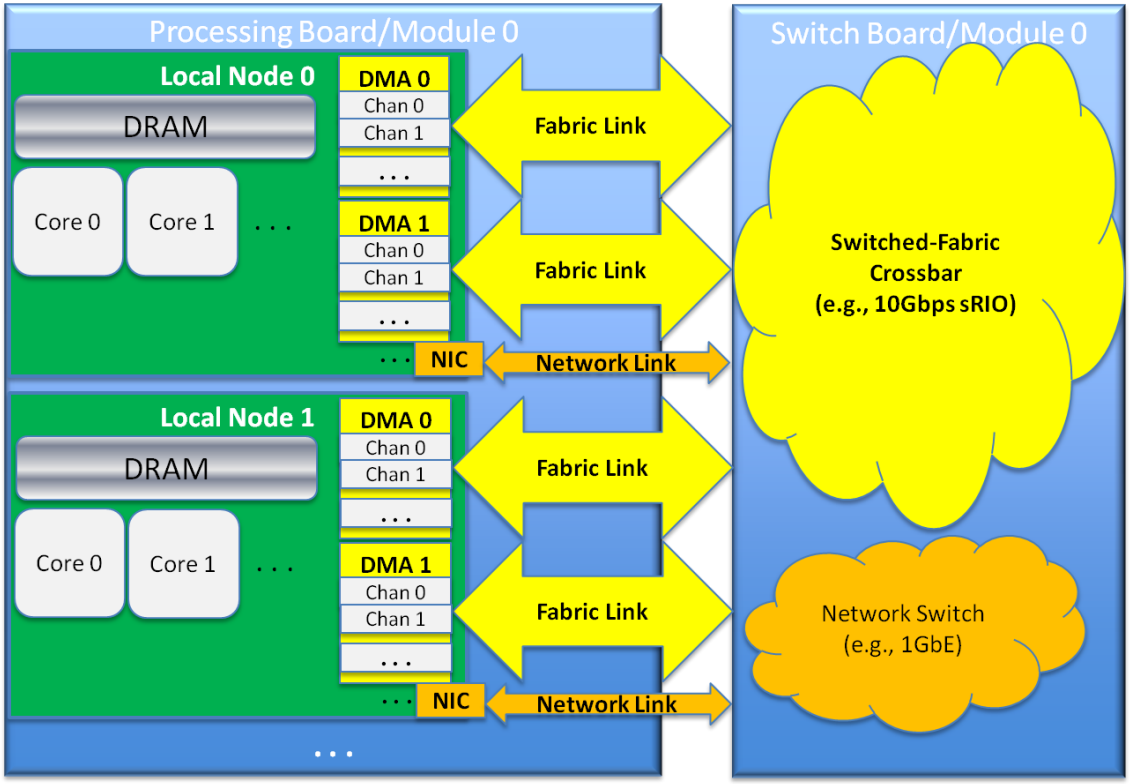
Connectivity – Data Plane

- Switched Fabric – High Performance Interconnect
- Examples: sRIO, Ethernet, IB, PCIe
- Other Options: Distributed Daisy-Chained / Mesh

OFED for Mercury sRIO

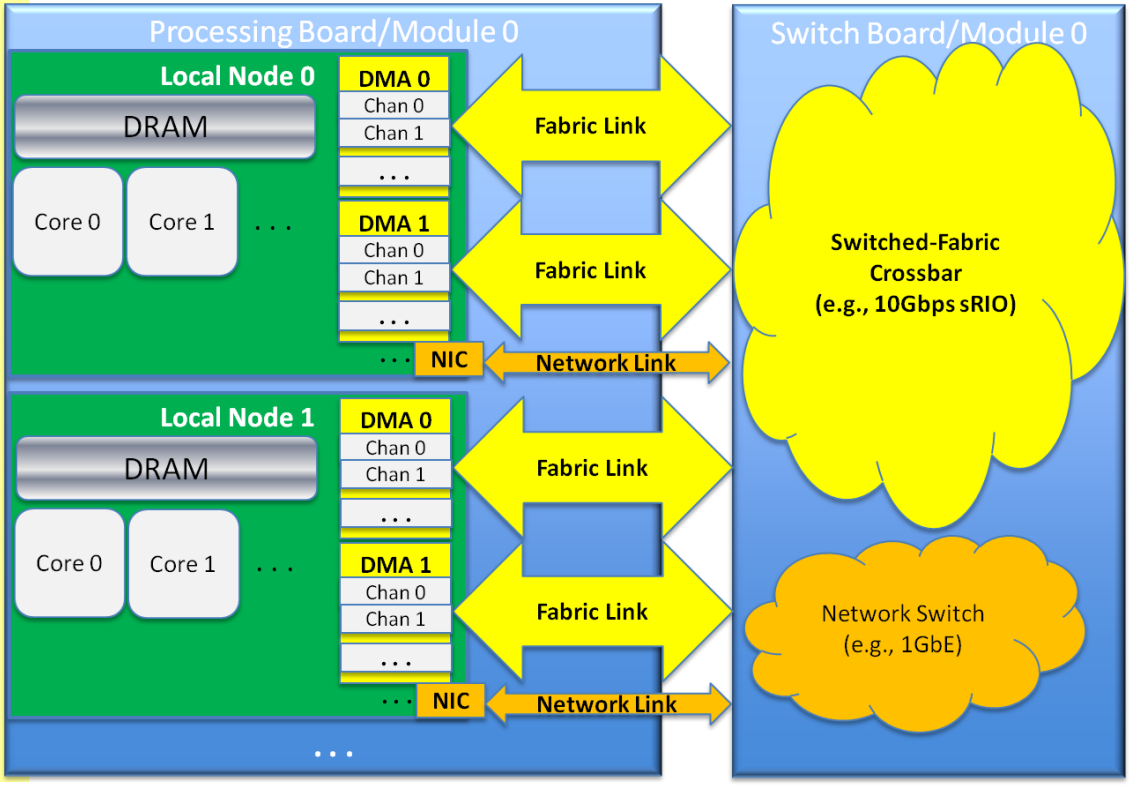
- CPU/Fabric HW Reference, Examples
- Mercury POET Engine
- Performance Data

CPU, Switch Board Model



Mercury Examples – CPU

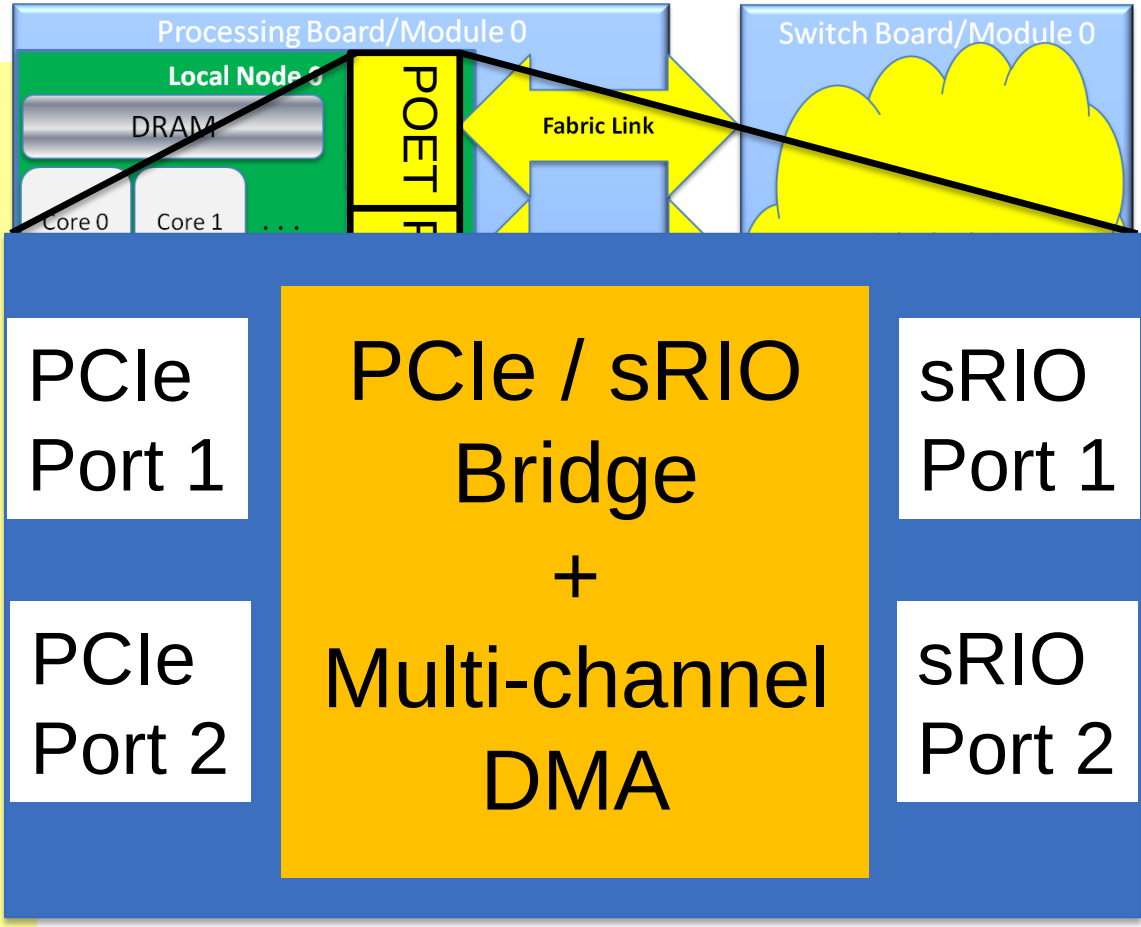
- Freescale 8641D dual-core
  - SoC sRIO DMA
- Core i7 dual core
- Xeon dual-quad core



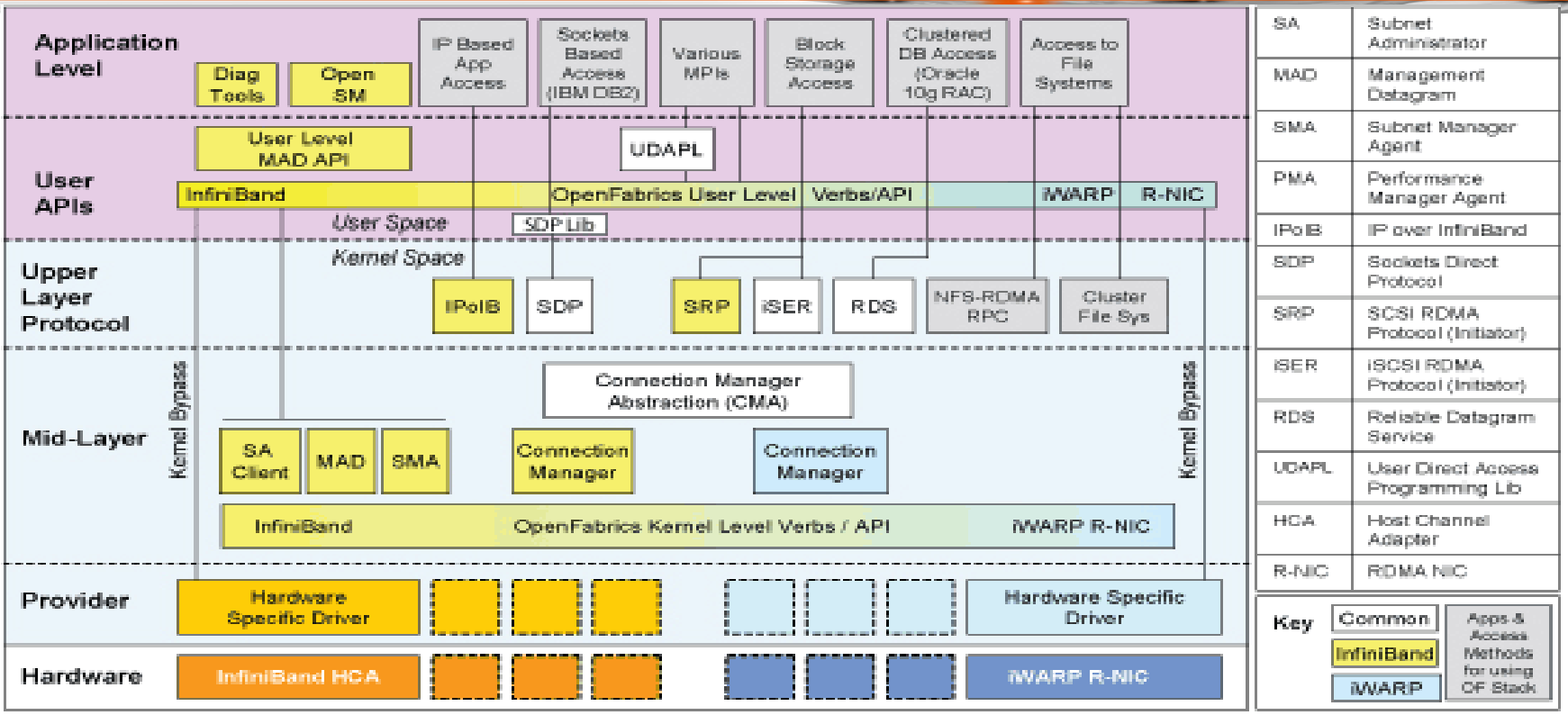
Mercury POET

Protocol Offload Engine Tech.

- Data Plane for Intel Embedded Nodes
- PCIe ←→sRIO DMA
- L2 Ethernet (Planned)



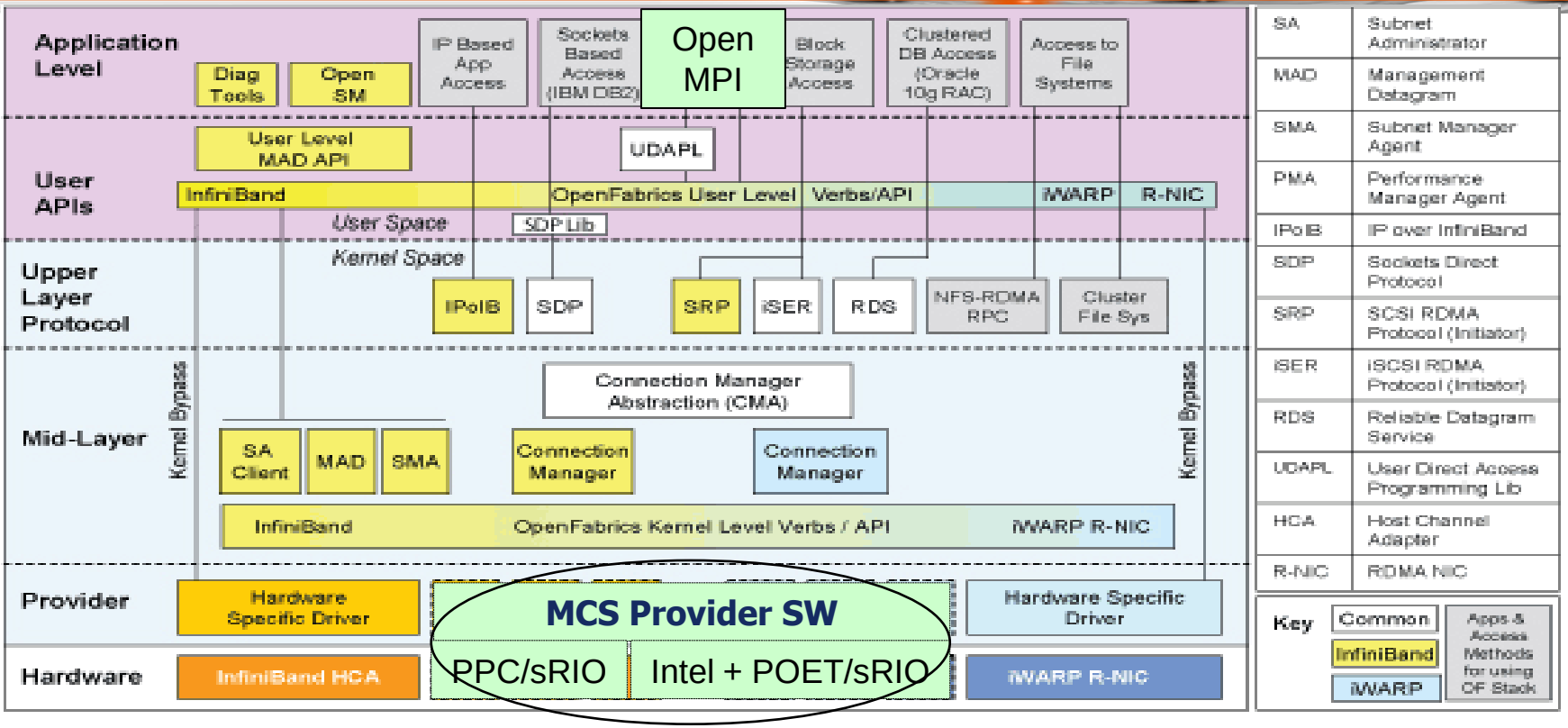
OFA Stack



SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA	Host Channel Adapter
R-NIC	RDMA NIC

Key: Common; InfiniBand; iWARP; Apps & Access Methods for using OF Stack

OFA Stack + Mercury Software

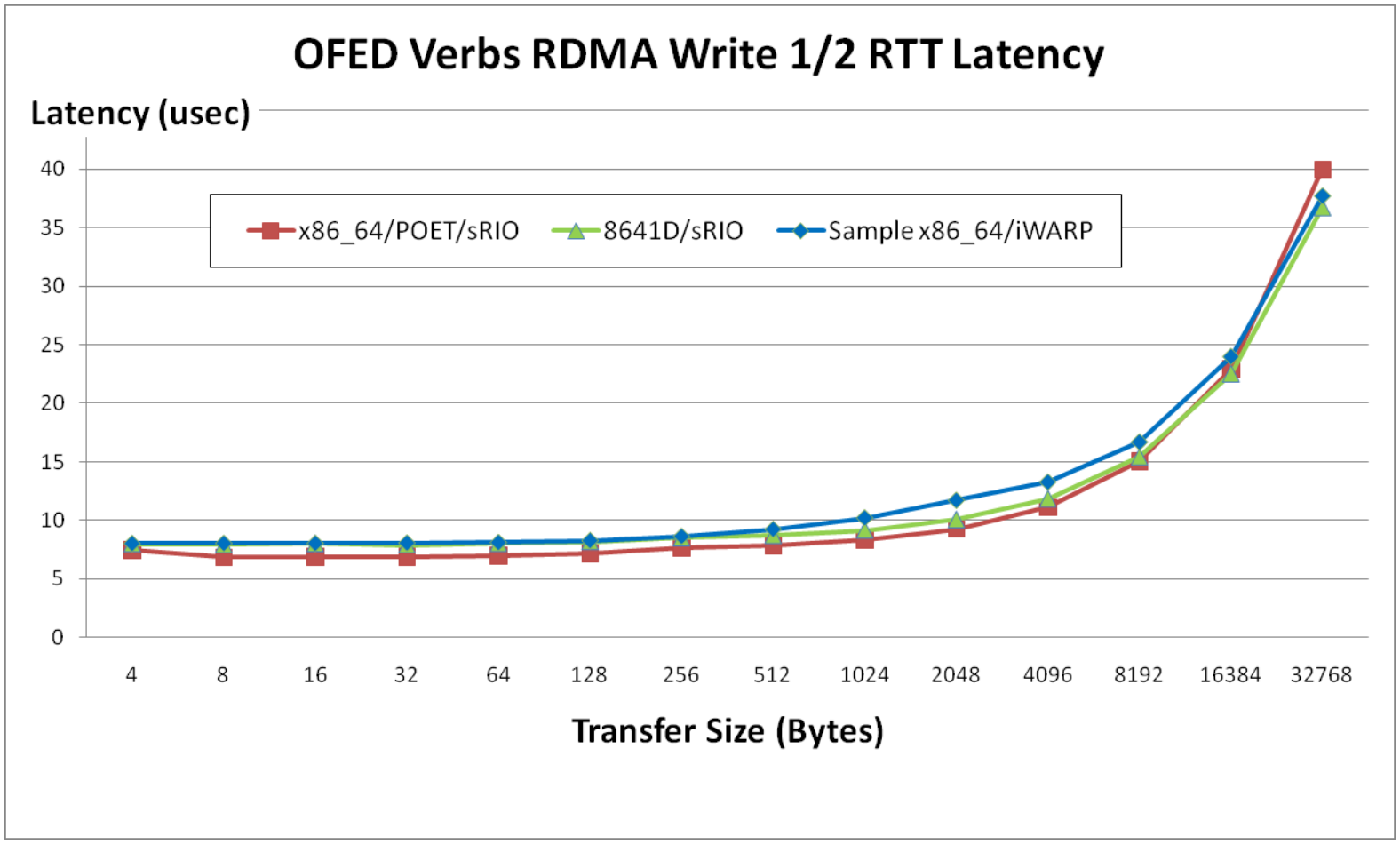


Performance Data Next

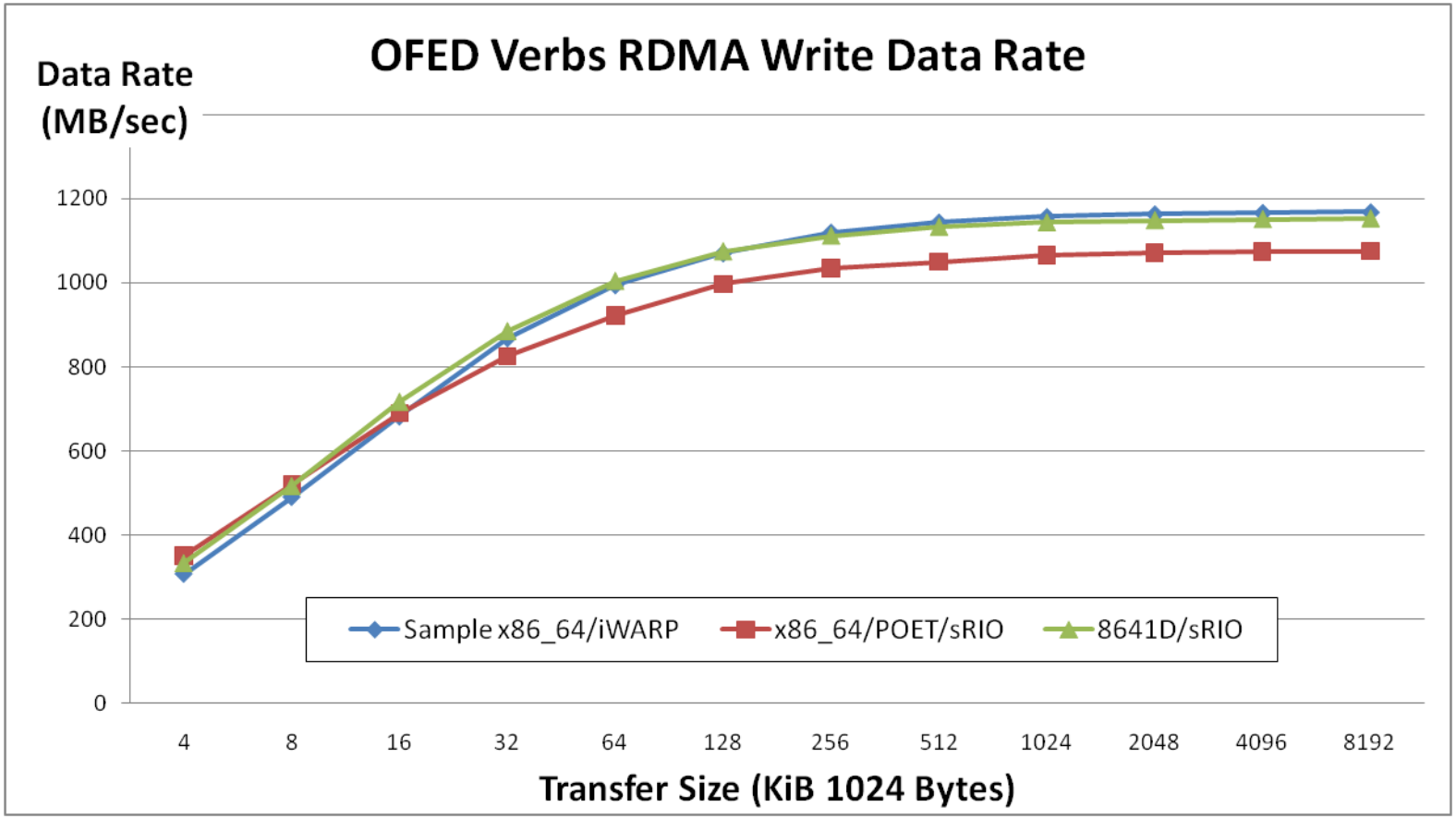
- Mercury Open MPI, OFED Software
  - Freescale 8641D / sRIO
  - Intel + POET/sRIO
- Sample 10GbE iWARP (same rate as sRIO)
- Mercury-developed Benchmark SW
  - Your mileage may vary

Intel/POET/sRIO Data is Preliminary

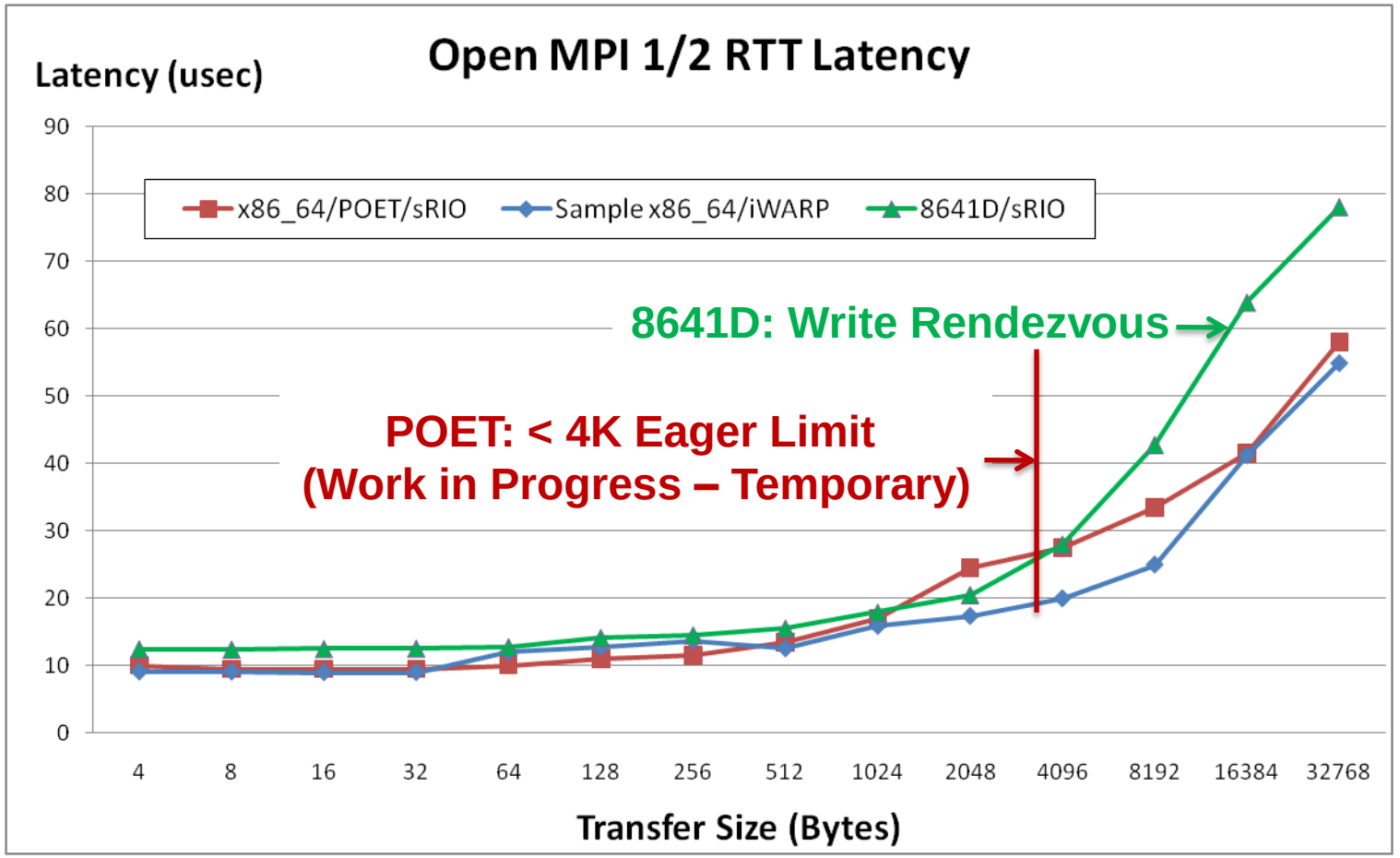
OFED Latency sRIO, iWARP



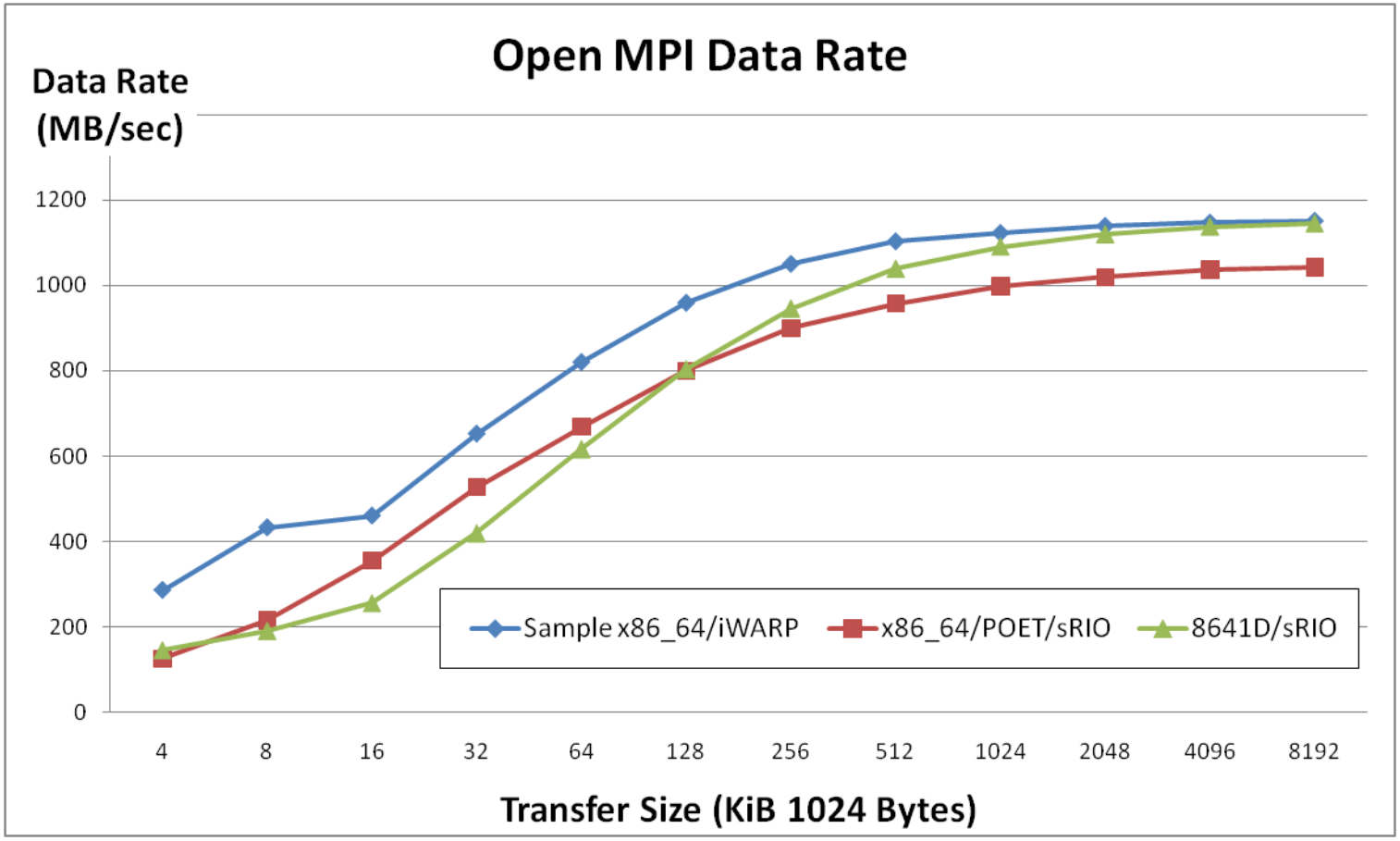
OFED Data Rate sRIO, iWARP



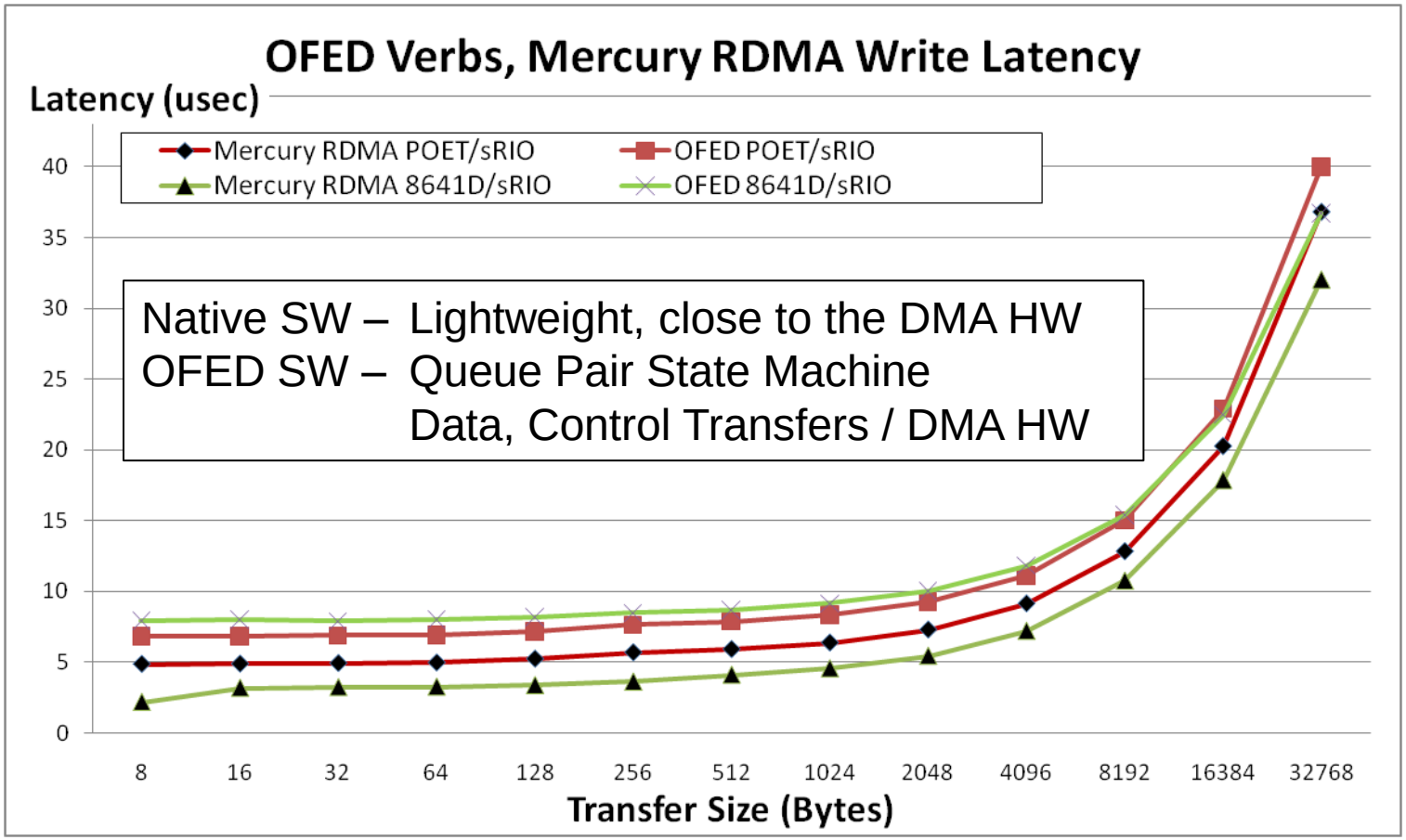
OMPI Latency sRIO, iWARP



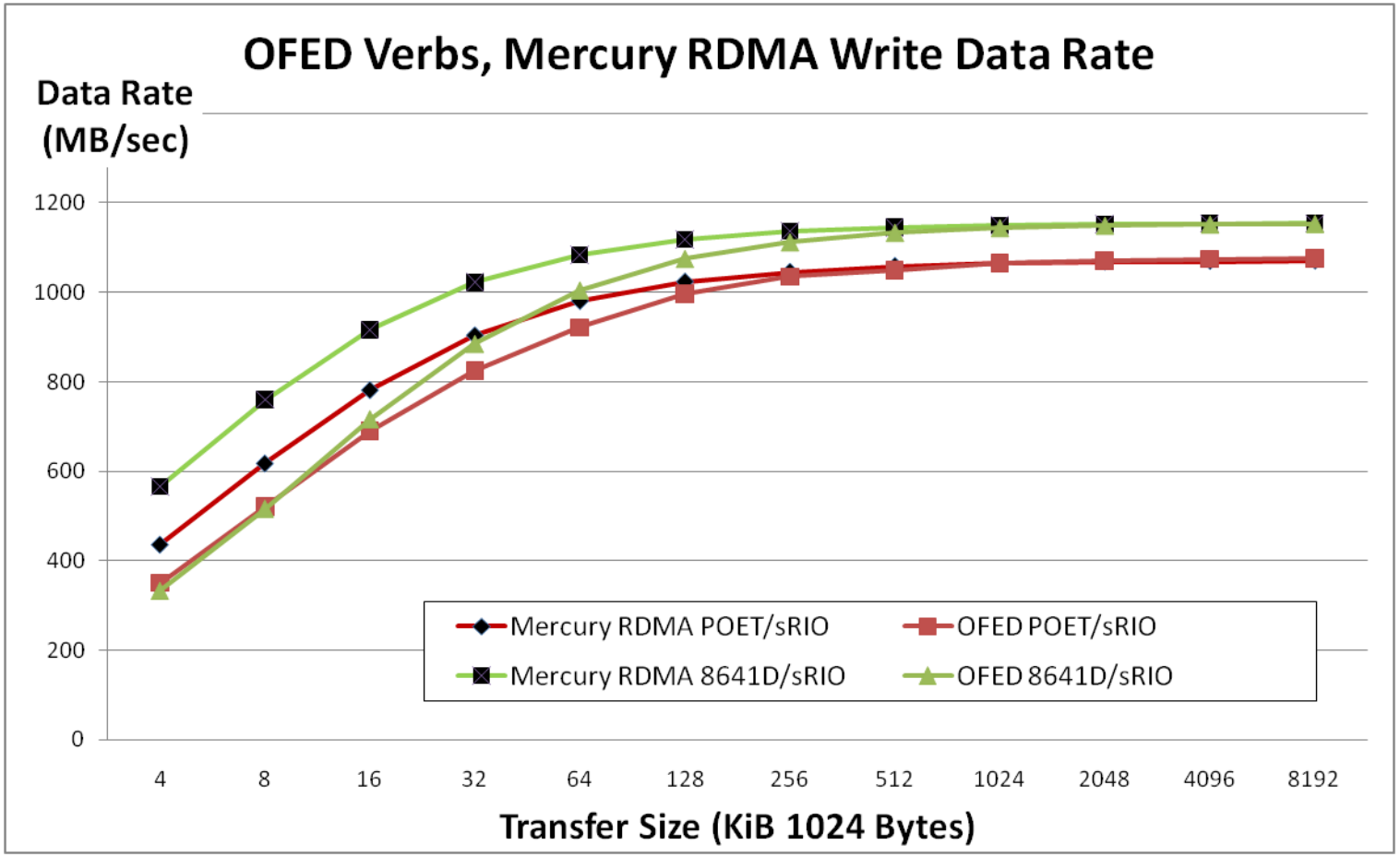
OMPI Data Rate sRIO, iWARP



OFED + Native Latency sRIO



OFED + Native Data Rate sRIO

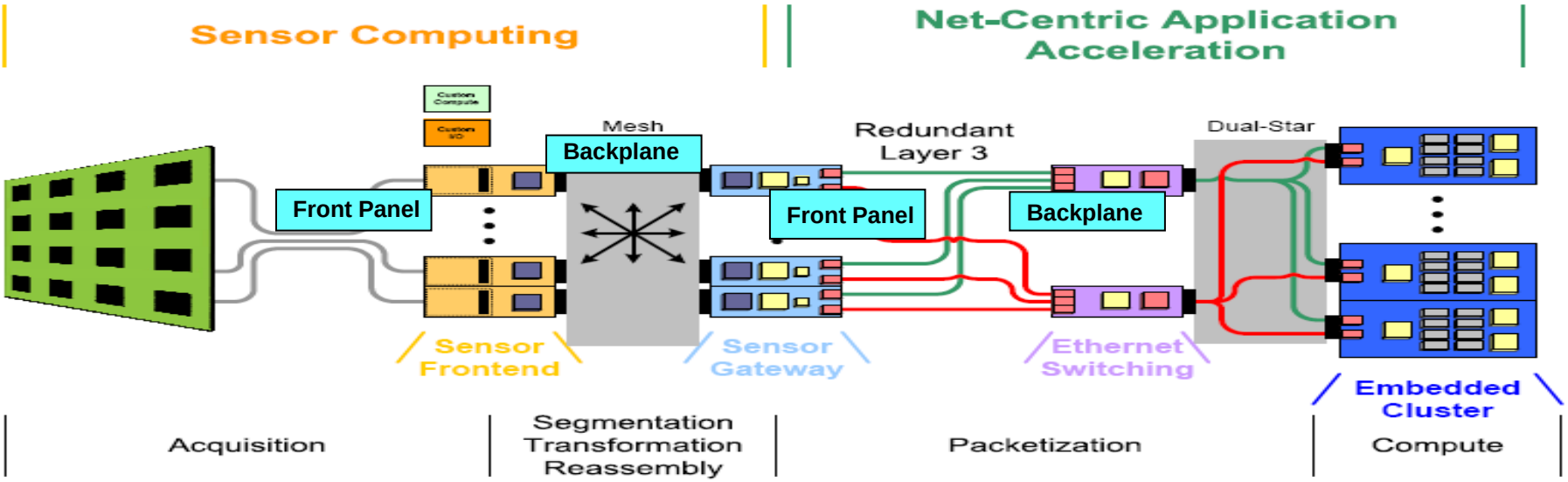


OPENFABRICS ALLIANCE

Embedded Data Flow Use Cases

- Fabric I/O – Memory Copies
- Sensor I/O – Streaming Mode
- Heterogeneous Compute Devices

Embedded Fabric IPC

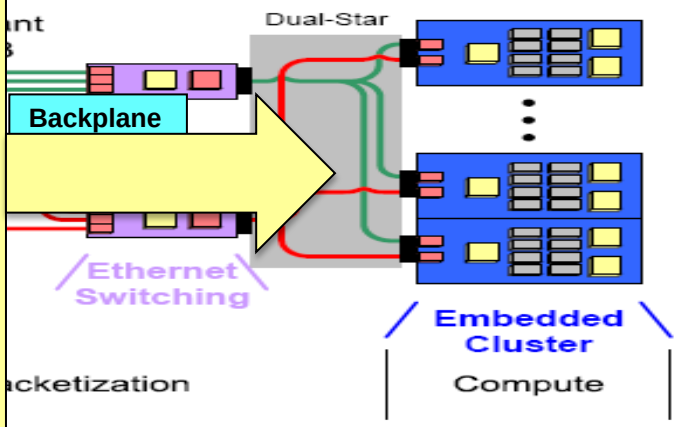


Embedded Fabric IPC

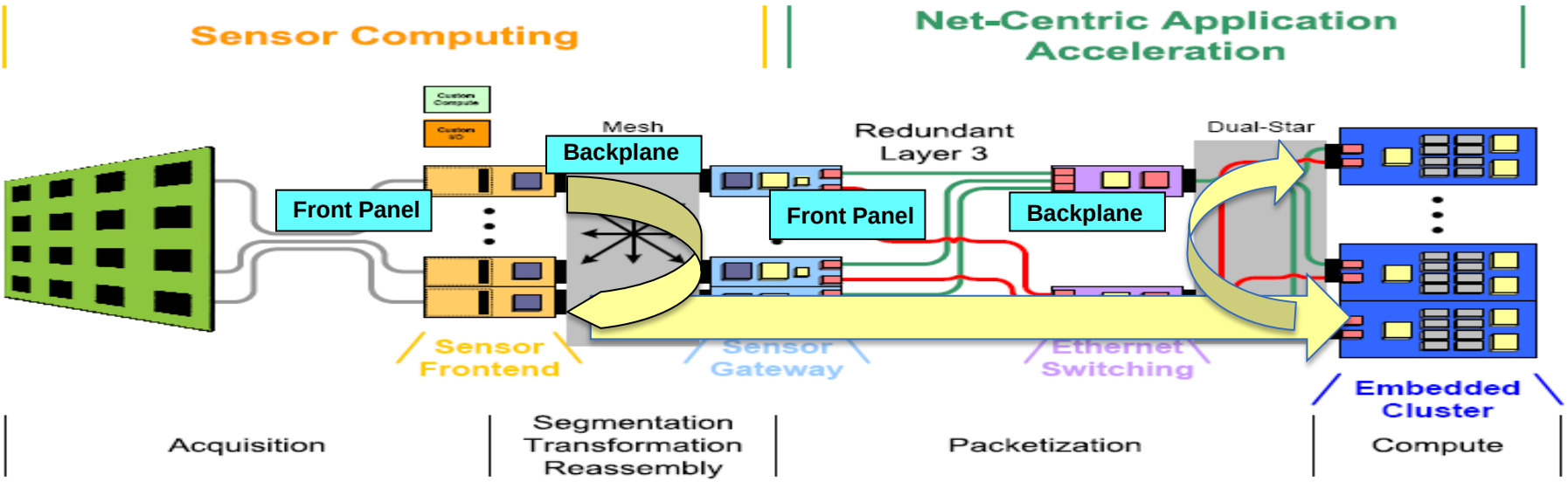
Sensor Computing

Net-Centric Application Acceleration

- Good Opportunity to use OFED
- Accelerators Present Challenges
  - Ex: FPGA direct connect to data plane
- Minimal Mem BW, CPU is Essential



Embedded Sensor I/O + IPC



- Sensor → Cluster: Need SW for IPC (or a Wire Protocol)
- Sensor Tx: Minimize CPU/SW Role
  - Ex: SW set up sensor IPC, then autonomous inner-loop
- Cluster Rx: Like Fabric IPC, Minimize Mem BW, CPU

OFED* Enhancement Opportunities

- * Software API Enhancements

Fabric IPC: HPC / Real-Time

*

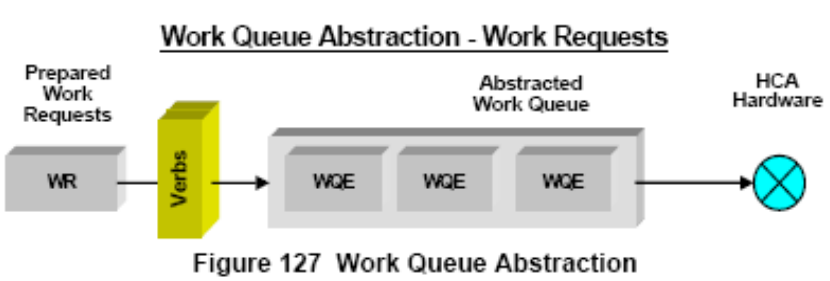


Figure 127 Work Queue Abstraction

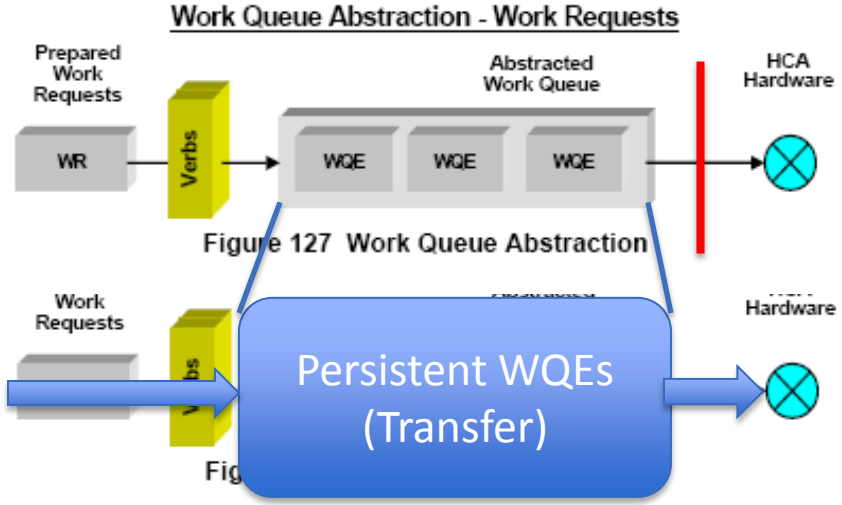


Figure 127 Work Queue Abstraction

Issue

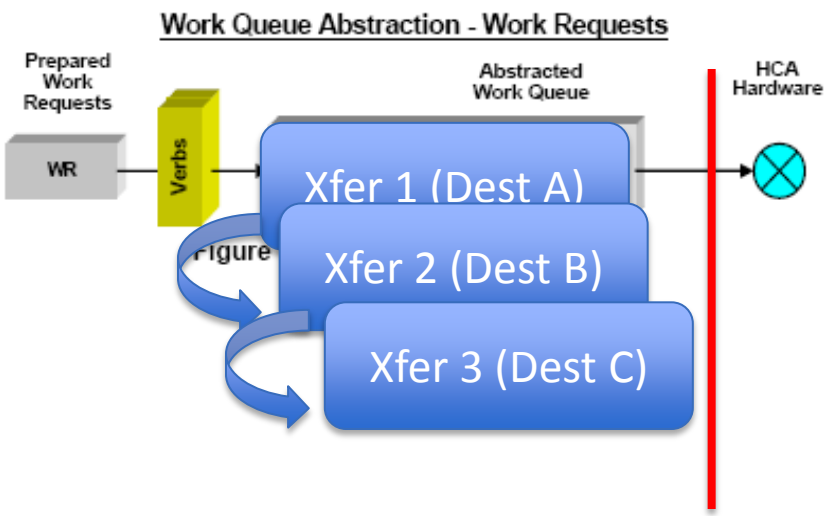
- Reinterpretation of WRs Each Time?

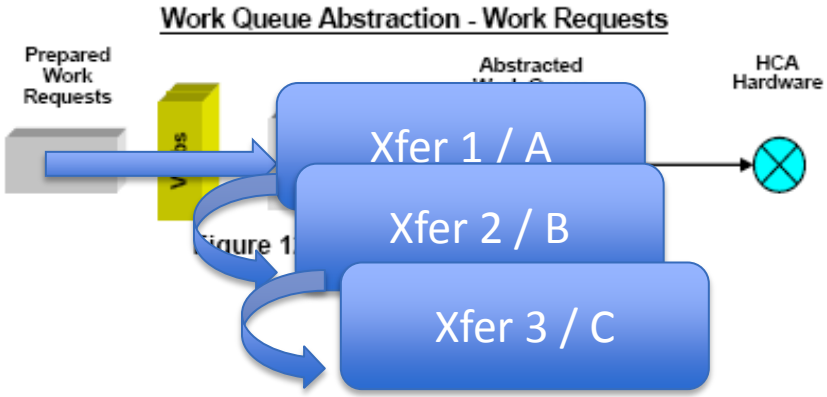
Improvement

- Pre-plan / Persistent
- Fast (Modify +) Queue

* Figure From InfiniBand specification

Now Multi-Node (Shared SQ?)





Outer Loop

- Setup Separate Transfers
  - One SW call per-destination
- Group Transfers into 1 Chain

Inner Loop

- Fast (Modify +) Queue

Contiguous / Physical Memory

- Contiguous Memory Benefits DMA Performance
  - PPC Ex: Block Address Translation (BAT) Mappings
- External Driver Provides non-Linux Managed Pool
  - MPI exposes to application via `MPI_Alloc_mem()`
- OFED Memory Registration Challenge
  - Patch libibverbs / avoid Linux `madvise()` issues
  - MCS provider distinguishes Linux / Non-Linux MEM
- Common Approach Too for Accelerator Memory?

OPENFABRICS ALLIANCE

Conclusions

Perspective

- Embedded Market Demands Flexibility
  - It's not just about IB, iWARP, RoCE, sRIO, PCIe
  - It is about open standard SW and performance
- Cost of Open / Efficiency – Always a Factor
  - OFED RDMA performance – promising initial results
  - MPI model adds performance challenges
- Hybrid Data Plane Programming Environments
  - HPC / Exascale: MPI investment focus→MPI+PGAS
  - HPEC: Vendor RDMA→ MPI+RDMA-based middleware

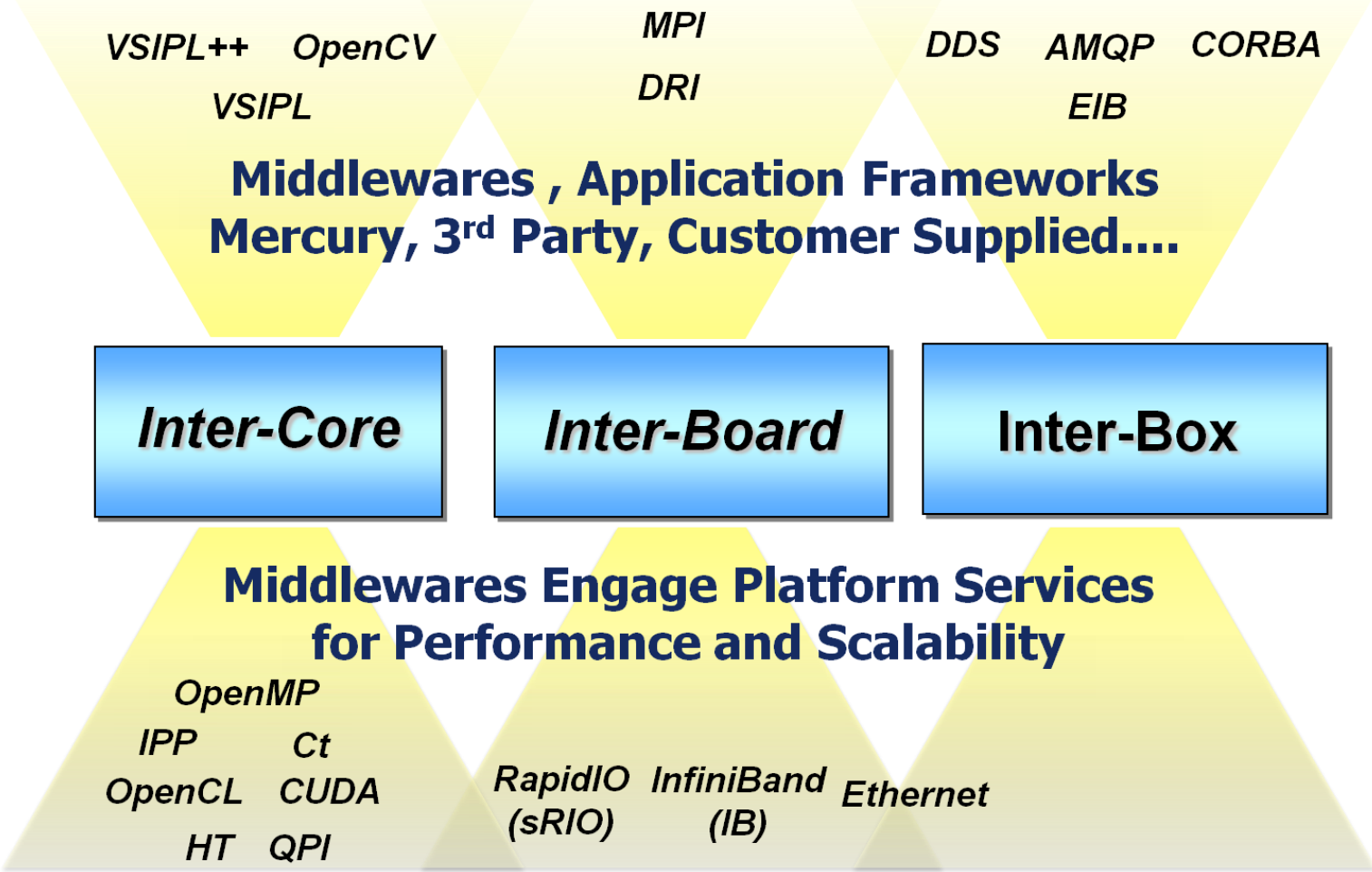
OPENFABRICS ALLIANCE

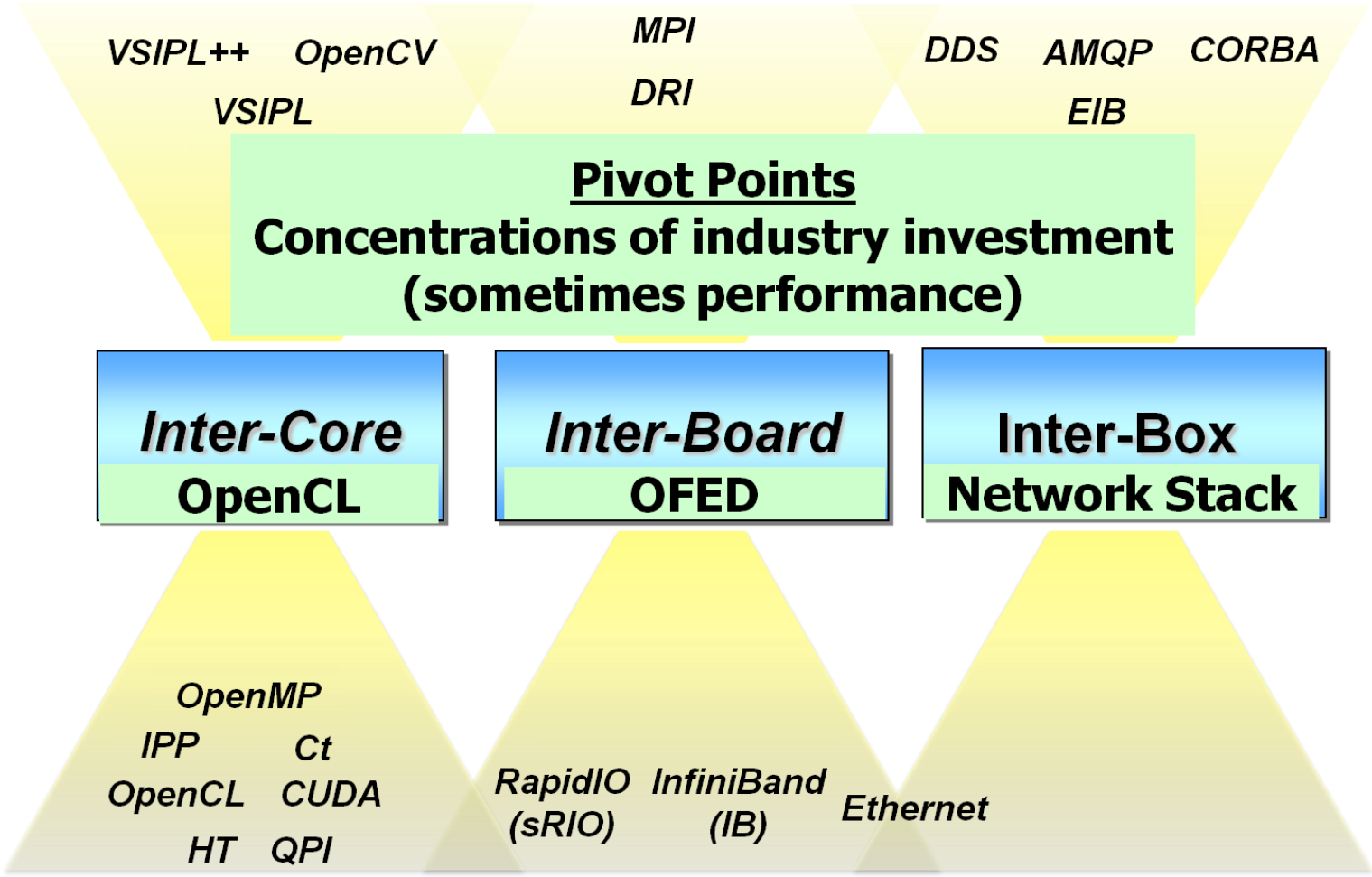
Backup

OPENFABRICS ALLIANCE

OFED Alignment with Embedded

- Architectural Flexibility
- Offload / RDMA
- Open Systems, Industry SW Ecosystem





Miscellaneous – (RDMA) CM

- Devices Mercury Has Integrated with OFED not:
  - IB
  - iWARP
  - Linux “netdev”
- Patched RDMA CM
  - Associate “names” (IP addr) with sRIO L2 addr
- RDMA CM Connection Support (for now):
  - Emulate “just enough” IB
  - RDMA CM layer over QP1 with send/recv verbs