Proximity-based Computing

OPENFABRICS ALLIANCE

David Cohen, Goldman Sachs

What is Proximity Computing

1. A business group uses rsync to replicate data from the intranet into a set of compute farms in advance of the execution of a job. These farms reside at various co-location facilities.

2. An App/Dev team sneakernets a couple hundred gigabytes to prime synchronization targets in London and Tokyo rather than using rsync to replicate the data from the NYC-based, master source.

3. An App/Dev team co-locates all of their software application clients (aka "clones") on the same network switch segment as the NAS filer where their shared data resides. Subsequently, the nightly batch job realizes a substantial reduction (more than 4x) in its time to completion.

4. Proprietary (Prop) and Automated Market Making (AMM) Trading functions migrate to co-location facilities at an Exchange

The distance between an application and its data has an impact on the application’s response time.

A few Principles

1. The Channel's Capacity (i.e. bandwidth) is largely a function of cost

"A single multiaccess computer would fill the bill if expense were no object, but there is no way, with a single computer and individual communication lines to several geographically separated consoles, to avoid paying an unwarrantedly large bill for transmission."
--- J.C.R. Licklider/Taylor, "The Computer as a Communication Device," 1968

2. Latency is a function of speed-of-light delay + transfer time

"if you have a network link with low bandwidth then it's an easy matter of putting several in parallel to make a combined link with higher bandwidth, but if you have a network link with bad latency then no amount of money can turn any number of them into a link with good latency."
--- Cheshire, "It's the Latency, Stupid!" 1996

3. The source of a message and controlling the transmission of that message are distinctly different problems.

"These semantic aspects of communication are irrelevant to the engineering problem."
--- Shannon, "A Mathematical Theory of Communication," 1948

4. The encoding of control bits happens during transmission

The principle, called the end-to-end argument, suggests that functions placed at low levels of a system may be redundant or of little value when compared with the cost of providing them at that low level.
--- Saltzer et al, "End-to-End Arguments in System Design," 1984

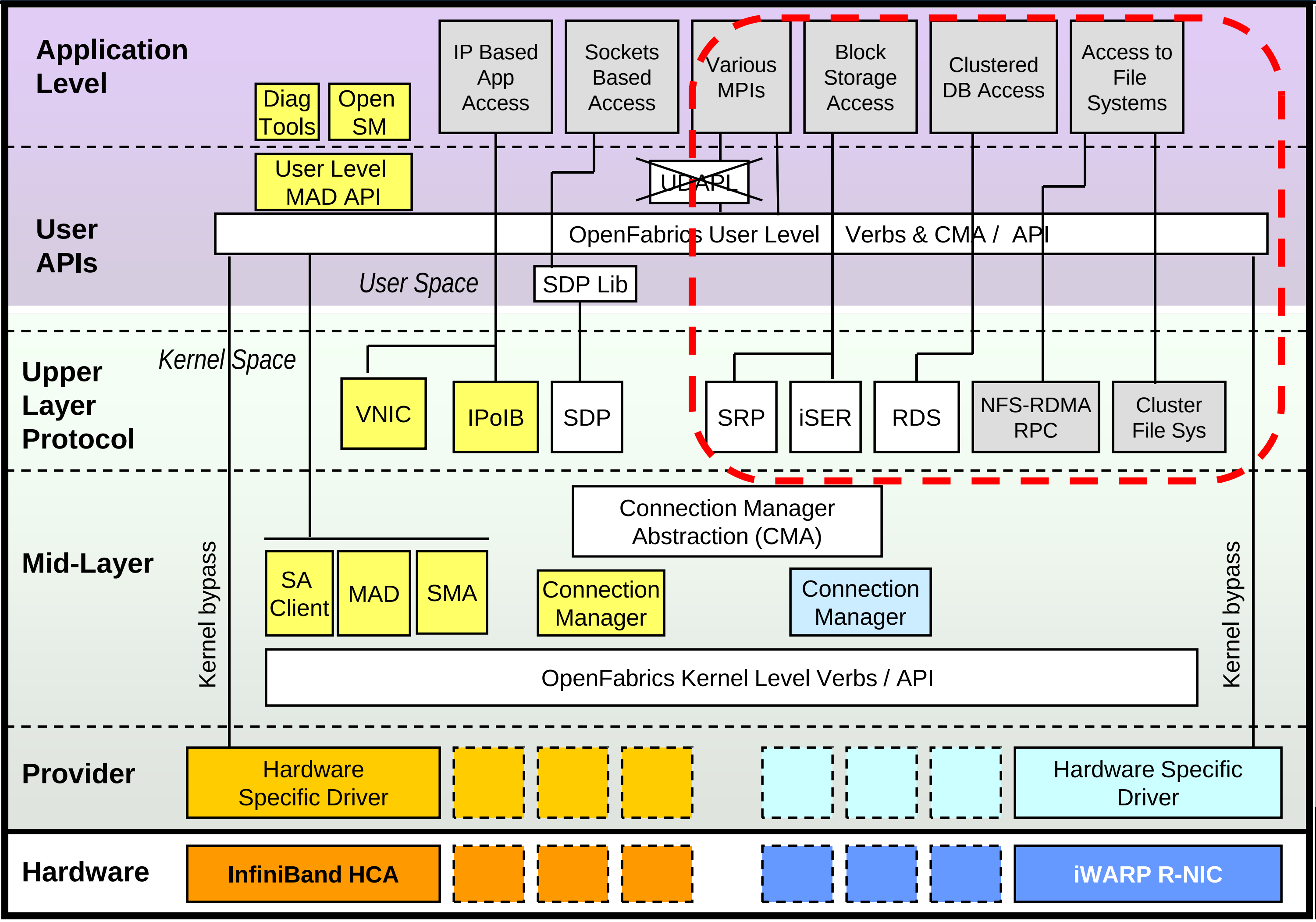
Bandwidth Delay Product (BDP)

The product of a Communication Channel’s data link capacity (bits-per-second) * its end-to-end delay (in seconds). This is a measure of data that has been transmitted but not yet received. In a reliable network, the sending host’s buffers are a function of the channel’s BDP.

> *As the channel capacity gets large (1Gbps and beyond) and the speed-of-light delay approaches zero, many people expect the impact of buffering to be reduced as well.*

The impact that buffering has on latency is a measure of the efficiency of the transport protocol being employed.

User Level Protocols (ULP) of Interest



Application Level

Diag Tools

Open SM

IP Based App Access

Sockets Based Access

Various MPIs

Block Storage Access

Clustered DB Access

Access to File Systems

User Level MAD API

~~UDAPL~~

User APIs

OpenFabrics User Level Verbs & CMA / API

User Space

SDP Lib

Kernel Space

Upper Layer Protocol

VNIC

IPoIB

SDP

SRP

iSER

RDS

NFS-RDMA RPC

Cluster File Sys

Mid-Layer

Kernel bypass

Connection Manager Abstraction (CMA)

SA Client

MAD

SMA

Connection Manager

Connection Manager

OpenFabrics Kernel Level Verbs / API

Kernel bypass

Provider

Hardware Specific Driver

Hardware Specific Driver

Hardware

InfiniBand HCA

iWARP R-NIC

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
VNIC	Virtual NIC
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

Platform Changes on the Horizon

1. Multi-core processors with NUMA memory design
   - AMD Shanghai and Hyper-transport (HT3)
   - Intel Nehalem and Quickpath (QPI)
2. The next two generations of PCI Express ("gen 2" and "gen 3")
3. I/O subsystem designed for Virtualization
   - Single Root I/O Virtualization (SR-IOV)
4. Storage Devices
   - A SATA-II disk drive increases in capacity to 2TB (and beyond); disk access operations get slower do to increase in seek times.
   - A Solid State device (SSD)/flash decreases data access (read and write) times by orders of magnitude.
5. Host Networking
   - Host comes with a dual port 10Gbps LAN-on-Motherboard (LOM) part
   - Small, inexpensive 4 to 48 port 10Gbps Ethernet switches come into the market
   - Ethernet is being enhanced to support higher levels of service (e.g.. Flow Control, Congestion Notification, etc)

Why not Infiniband?

1. Enterprise Data Centers are dominated by Ethernet
   - Large scale Infiniband fabrics are not feasible (i.e. more then 30 to 50 severs)
     - Transparency, Debugging, etc are consistently brought up as concerns
   - Network Engineers have shown a willingness to consider Infiniband within the context of a small link layer domain (i.e. must fit within a single data center rack)
2. Since most Business Plans call for growing port count/density at the core there is a mismatch with Enterprise opportunities
3. But…
   - A link layer segment is often used to isolate a set of applications from the broader, general purpose constituency of applications. We refer to this isolated segment as a Pod. A Pod is one or perhaps as many as three racks of computers in a single link layer domain.
   - Like Fibre Channel, Infiniband's network and transport layers can be encapsulated within an Ethernet frame.

What about iWARP?

1. We should drop the use of the acronym “iWARP”
   - RDDP/DDP plus a transport adaptation (e.g. SCTP and MPA/TCP)
   - RDMAP/DDP/MPA/TCP is what people are referring to when the say iWARP.
2. RDMAP/DDP/MPA/TCP (along with TOE) has failed to gain traction relative to Infiniband
   - Part of the inertia has been reliance on TOE as its firmware implementation is considered by many to be non-scalable.
   - TOE is also a point-in-time solution that must be upgraded.
   - Even if RDMAP/DDP/MPA/TCP was decoupled from TOE its not clear things will get better
3. As the “Proximity” of participants in a communication increases the use of IP-based Protocols has limited value relative to the overhead.
   - When two applications that exchange messages are co-located in a Pod the response time expectation is different when their communication traverses one or more routers. (see Principles and BDP slides)
4. The following slides touch on some of the issues to be considered

Converged Enhanced Ethernet (CEE)

- 802.1Qau - Congestion Notification
  - Support for higher layer protocols that are highly loss or latency sensitive
  - VLAN tag encoded priority values
- 802.1Qaz - Enhanced Transmission Selection
  - Share bandwidth between priorities
- 802.1Qbb - Priority-based Flow Control
  - Delivery of data frames without frame loss due to congestion
  - Similar to Pause but operates on VLAN tag priority values

CEE allows us to reconsider Ethernet as a lossless fabric.

Host demand for more than 1Gbs of bandwidth is increasing

- multi-socket motherboard
- multicore processors
- increased memory and I/O bandwidth
- increasing use of resource partitioning techniques such as system virtualization

Latency decreases as distance between hosts decreases. At the same time, bandwidth increases are enabling very high throughput data transfers between hosts. The result is that the cost/overhead of network related processing at the hosts is increasing.

The Efficiency of the Network and Transport Protocols is becoming more important!

Adoption of Hybrid Storage

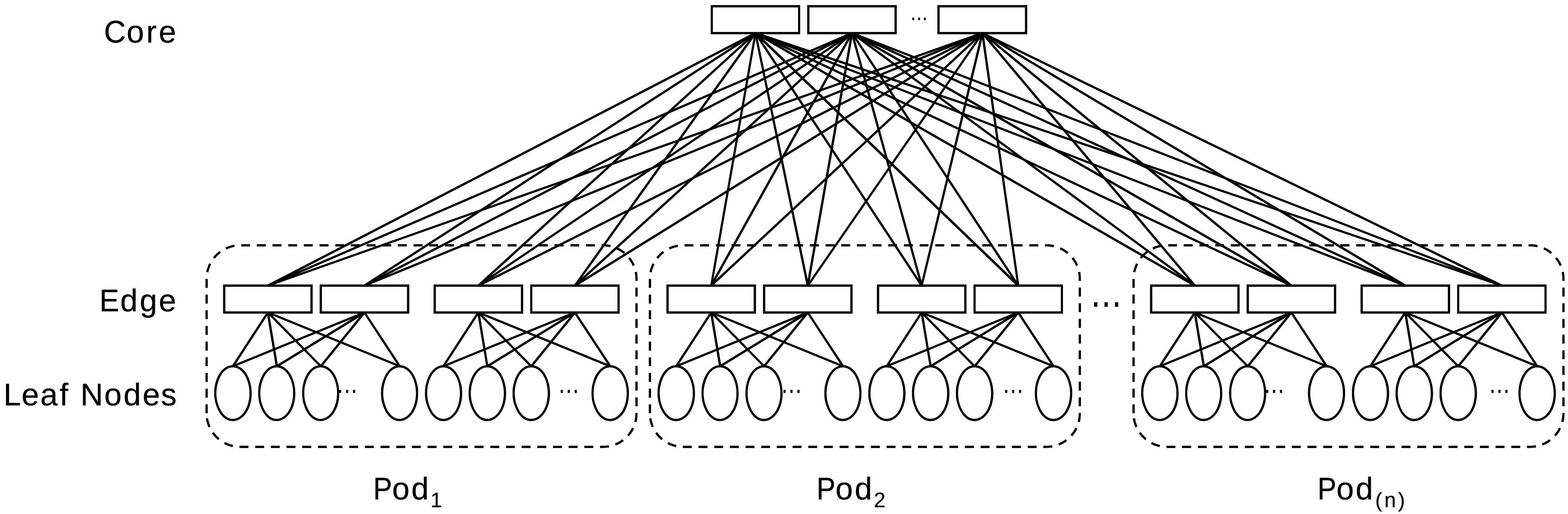
- Combine SATA for capacity and SSD/Flash for performance
- Transparent data placement using virtual block device and/or virtual memory techniques.
- Sun's ZFS and NetApp's WAFL file systems are examples.

Increasing demand for very fast access to storage is placing a lot of focus on caching techniques. This is similar to WAN Optimizations that occurred several years ago. See dm-cache, NFS-Ganesha, fs-cache, varnish, memcached, etc.

“In other words, putting these things on the fabric changes the fabric requirements!”

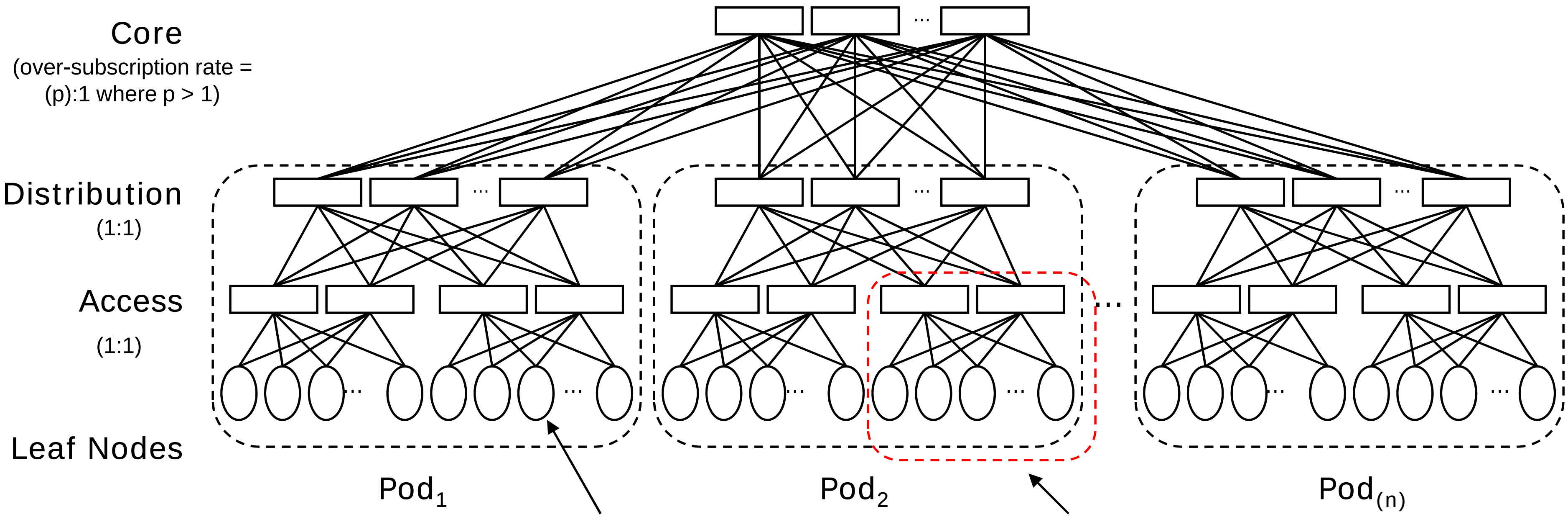
Does the Core-Edge Data Center Network Topology Make Sense Going Forward?

Traditional Core/Edge Topology



http://ccr.sigcomm.org/online/files/p63-alfares.pdf

New Clos/Fat-tree Topology Using CEE



A node is a virtualized system

This function is provided by the NIC

http://ccr.sigcomm.org/online/files/p63-alfares.pdf

I/O Virtualization becomes mainstream

- Memory isolation/protection between virtual systems via IOMMU
- PCIe's SR-IOV plus hardware assists such as IOMMU.
- SR-IOV's Virtual Function (VF) extends the I/O bus to guest/containers.

Each VF is in essence a virtualized PHY/MAC. The result is that the host's I/O adaptor is being transformed into an L2 switch. Given the Clos/Fat-Tree topology above and the Pod architecture the host's adaptor becomes the Pod's access layer.

Pushing the Access Layer down to the Host's NIC

Assume that the access layer switches in a Pod are implemented on a node's NIC and that the NIC is PCIe/SR-IOV enabled.

- Sun's Crossbow
- Cisco/Vmware VMDirectPath + Virtualized Switches
- HP's Virtual Connect
- Others (e.g. Xsigo, 3-Leaf, etc)

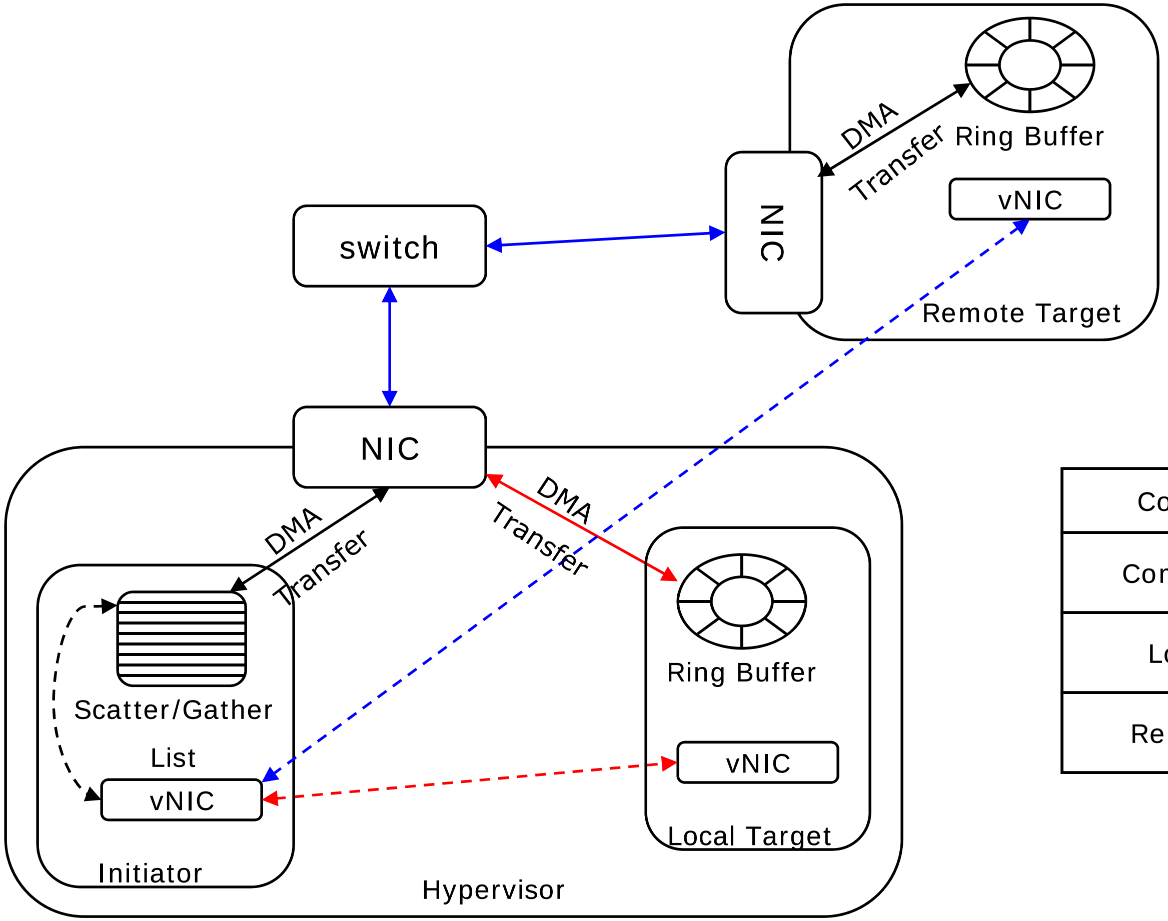
Dong et al, "SR-IOV Networking in Xen: Architecture, Design and Implementation," 2008
http://www.usenix.org/events/wiov08/tech/full_papers/dong/dong.pdf

Santos et al, "Taming Heterogeneous NIC Capabilities for I/O Virtualization," 2008
http://www.usenix.org/events/wiov08/tech/full_papers/santos/santos.pdf

Tripathi et al, "OpenSolaris Project
Crossbow: Network Virtualization & Resource Partitioning," 2008
http://opensolaris.org/os/project/crossbow/Docs/Crossbow_WP.pdf

Russell, "virtio: towards a de-facto standard for virtual I/O devices," 2008
http://portal.acm.org/ft_gateway.cfm?id=1400108&type=pdf

Support for Local and Remote RDMA

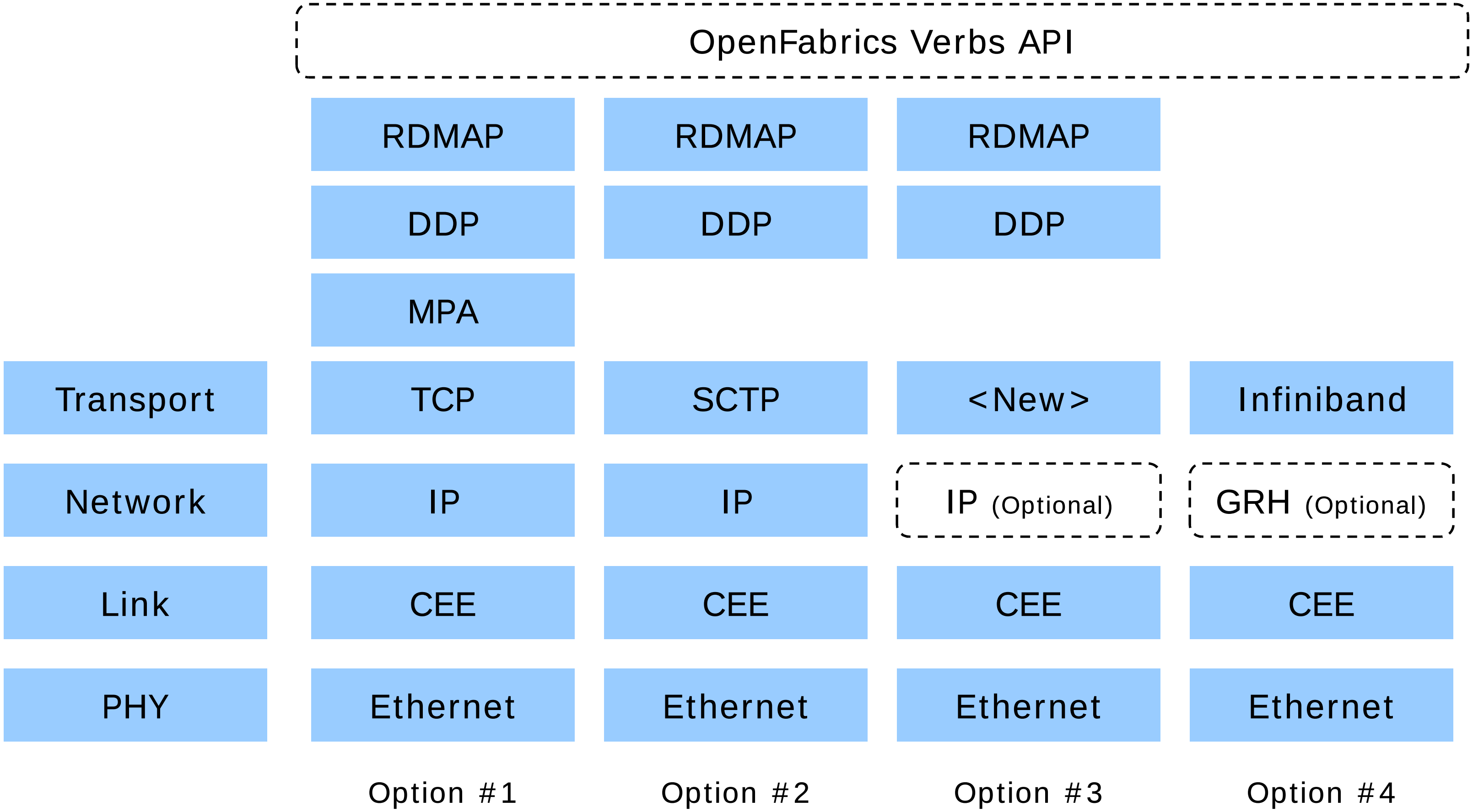


Communications Legend	
Common	——— - - - - - - -
Local	——— - - - - - - -
Remote	——— - - - - - - -

Where can we expect optimizations to occur?

1. Head-of-Line (HOL) Blocking
   - VLAN tag based priority values to break out traffic; TCP falls in the "Best-Effort" queue
   - This means ULP control bit settings must be scrutinized to insure efficient classification.
2. Network and Transport Protocol overheads
   - As Bandwidth Delay Product (BDP) approaches zero the overhead of buffer management becomes a huge percentage of the total latency
   - Requires high-frequency RTT measurements
   - Host network buffers in IP protocols don't go away
3. Within a Pod, employ RDMA for Storage Protocols
   - NFS/RDMA
   - iSER

OFA Options for Supporting CEE



OpenFabrics Verbs API

	RDMAP	RDMAP	RDMAP	
	DDP	DDP	DDP	
	MPA			
Transport	TCP	SCTP	<New>	Infiniband
Network	IP	IP	IP (Optional)	GRH (Optional)
Link	CEE	CEE	CEE	CEE
PHY	Ethernet	Ethernet	Ethernet	Ethernet
	Option #1	Option #2	Option #3	Option #4

Conclusions

- Proximity Computing
  - Going forward, the placement of data and the applications that consume it becomes increasingly important.
- The composition of a data center shifts toward Pods
  - The network optimization function simultaneously solves for *"maximize throughput and minimize latency while protecting against long-tail latency distributions (a.k.a. 'Jitter')"*
- Communication within a Pod drives the need for a transport protocol that is optimized for RDMA
  - As the bandwidth delay product approaches zero (i.e. bandwidth greater than 1Gbps and latency less then 100us) the cost of buffering becomes increasingly expensive.
  - We believe these factors weight heavily in favor of a Transport Protocol that can layer directly over CEE when constrained to a Pod.