Networks & Storage

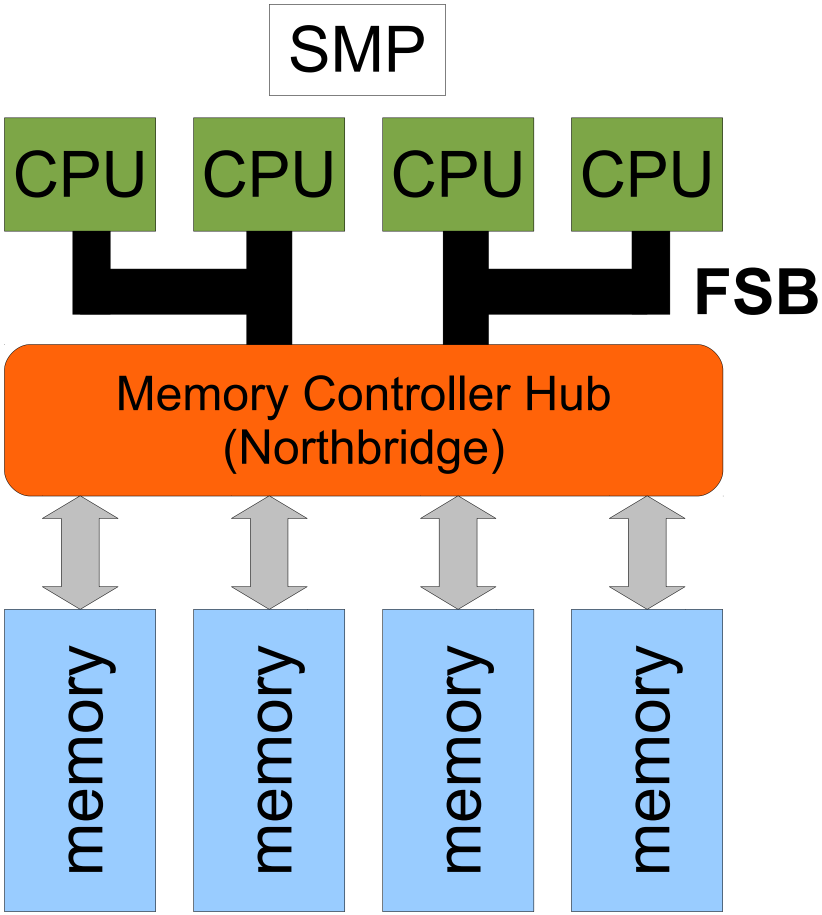
University of BRISTOL

Overview

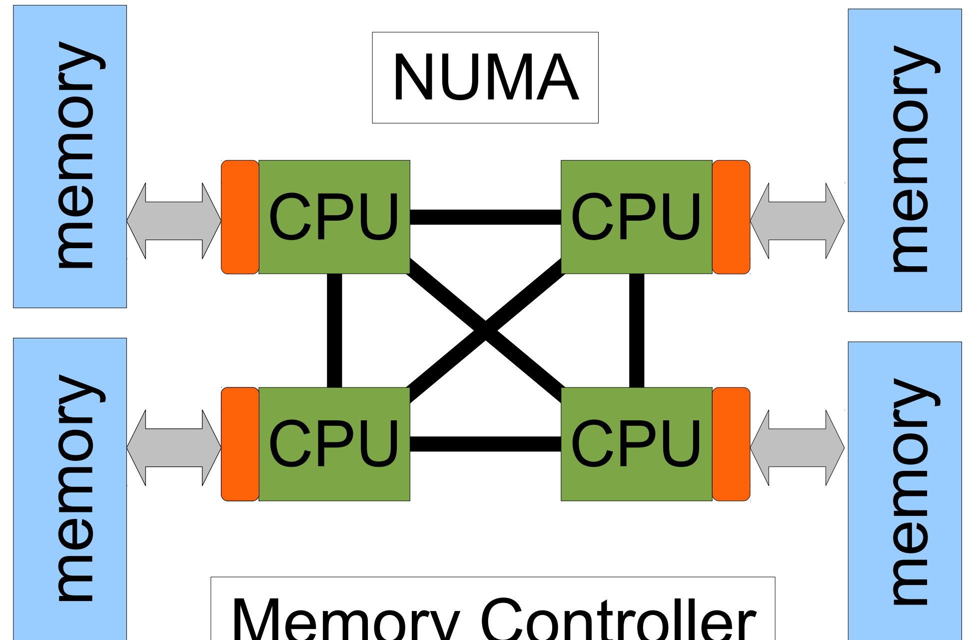
- Networking from the inside out
- SMP vs NUMA
- TCP/IP, Infiniband, RoCE, iWARP
- Performance of current systems
- Topologies
- Future trends
- Storage – Parallel File-systems vs. single disk
- Summary

University of BRISTOL

Inside the box: SMP vs. NUMA



FSB: Front Side Bus



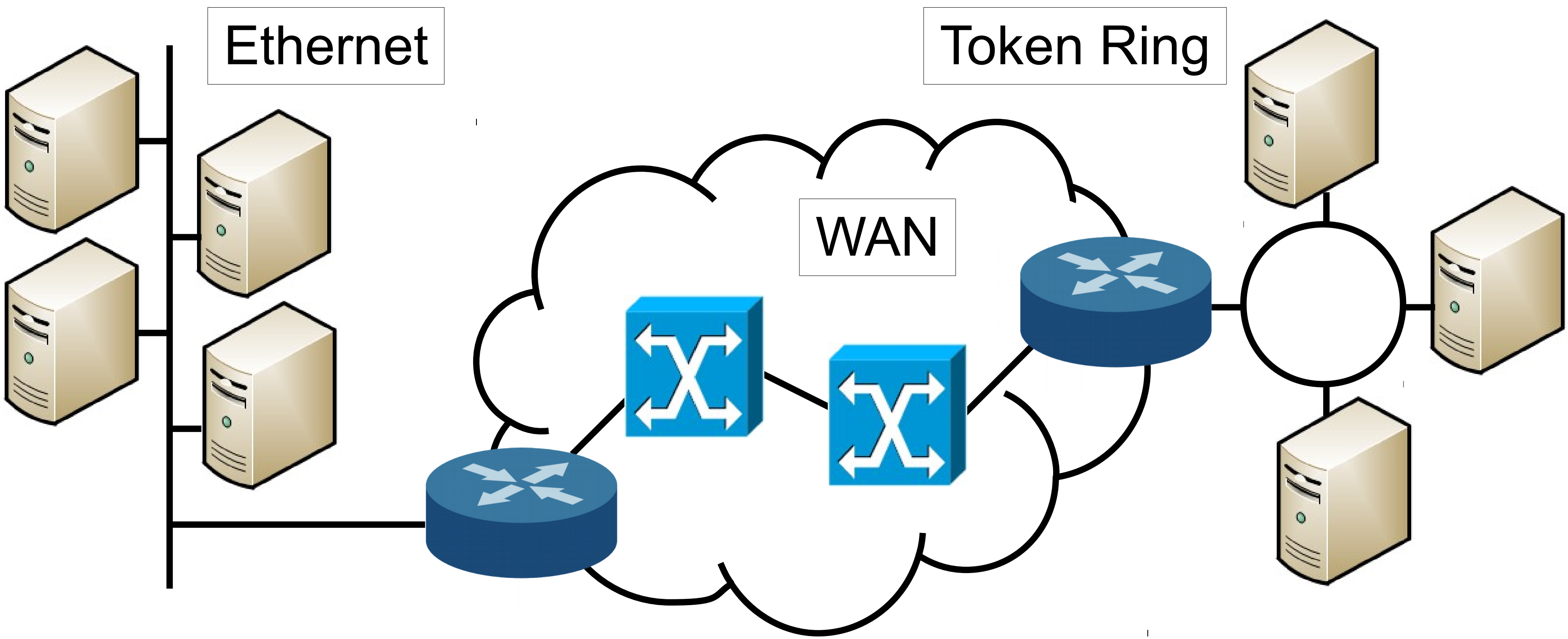
Memory Controller moved onto chip

- QuickPath (Intel) &
- HyperTransport (AMD)

are NUMA technologies

University of BRISTOL

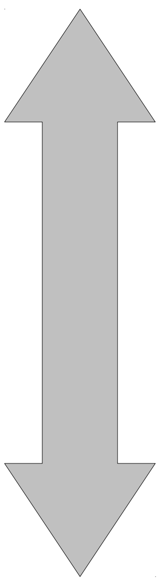
History: Network of Networks



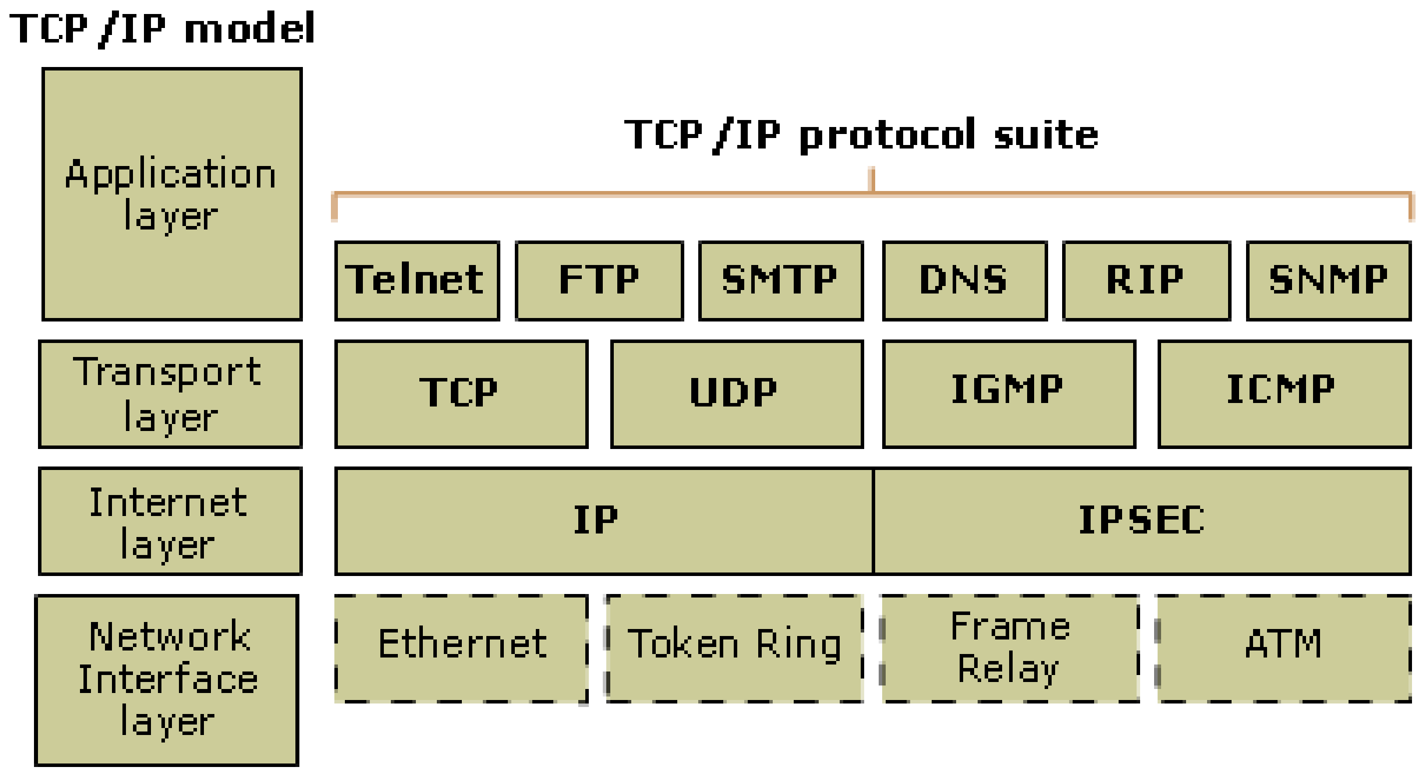
- TCP/IP US DoD project to connect networks from different vendors
- Designed to be fault tolerant
- IP: Move packets from node to node (4 byte address)
- TCP: Responsible for safe delivery (can trigger retransmission)

University of BRISTOL

TCP/IP



Buffers copied



Heavily Standardised

University of BRISTOL

Example: Traditional TCP/IP Networking Stack

User Application	Web browser
Application	HTTP
Transport	TCP
Network	IP
Link	Ethernet driver
Hardware	Ethernet

University of BRISTOL

Sockets: Example Server

```
#!/usr/bin/env python

import socket

TCP_IP = '127.0.0.1'
TCP_PORT = 5005
BUFFER_SIZE = 1024

s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
s.bind((TCP_IP, TCP_PORT))
s.listen(1)

conn, addr = s.accept()
print 'Connection address:', addr
while 1:
    data = conn.recv(BUFFER_SIZE)
    if not data: break
    print "received data:", data
    conn.send(data)  # echo
conn.close()
```

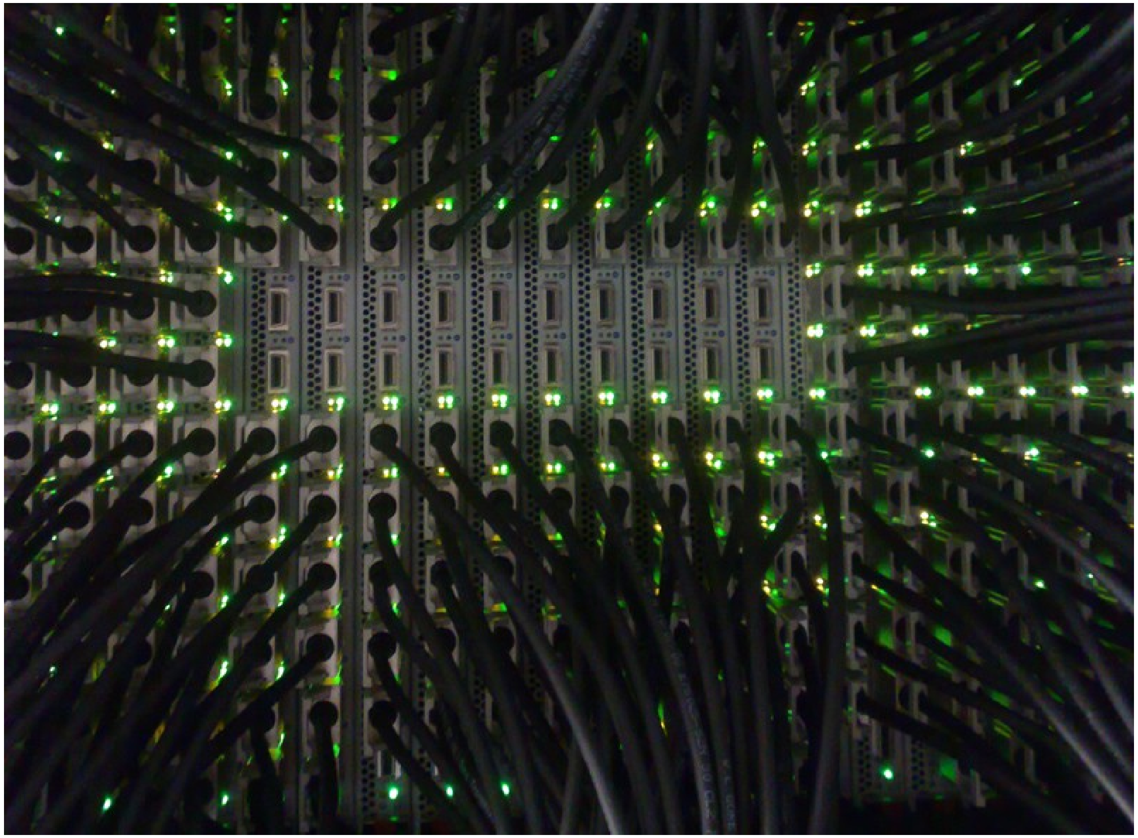
Sockets: Example Client

```
#!/usr/bin/env python

import socket

TCP_IP = '127.0.0.1'
TCP_PORT = 5005
BUFFER_SIZE = 1024
MESSAGE = "Hello, World!"

s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
s.connect((TCP_IP, TCP_PORT))
s.send(MESSAGE)
data = s.recv(BUFFER_SIZE)
s.close()

print "received data:", data
```

- Sockets provide a simple application interface
- But, we pay a price for copying buffers..

University of BRISTOL

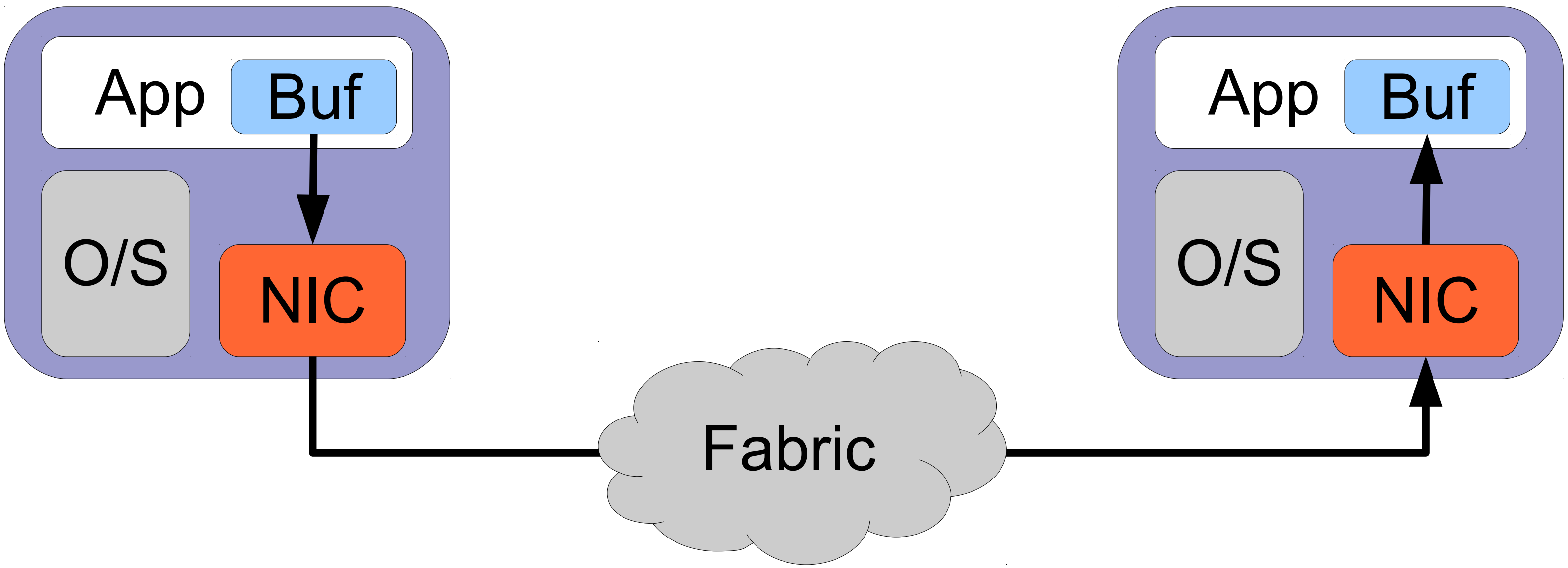
HPC & Networking



All well & good, but in HPC,
we want the fastest possible networking..

University of BRISTOL

Infiniband



Different design criteria to TCP/IP:

- A ***'messaging service'*** over a ***persistent channel***
- All hosts use Infiniband Fabric interconnect
- 'Stack bypass' – no O/S involvement
- Application can access messaging service directly
- Leads to lower latency communications

University of BRISTOL

Infiniband

Application

Infiniband Messaging Service

S/W Transport Interface

Transport

Network

Link

Physical

Queue Pairs

Send Recv

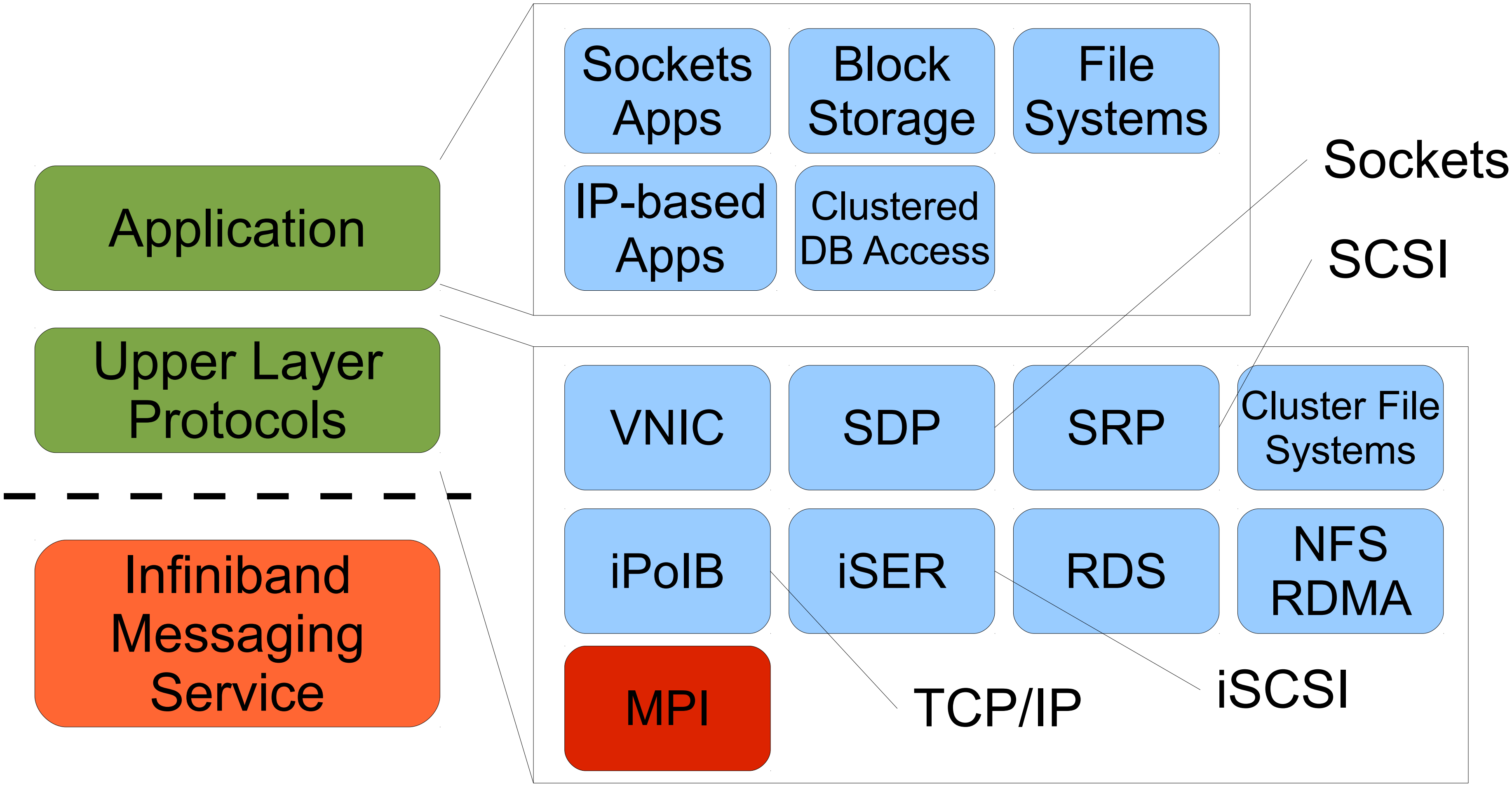
RDMA Message Transport Service

There is still an Infiniband network stack, but much is offloaded onto the Host Channel Adapter (HCA)

RDMA:
Remote
Direct
Memory
Access

University of BRISTOL

Infiniband: Many provided protocols



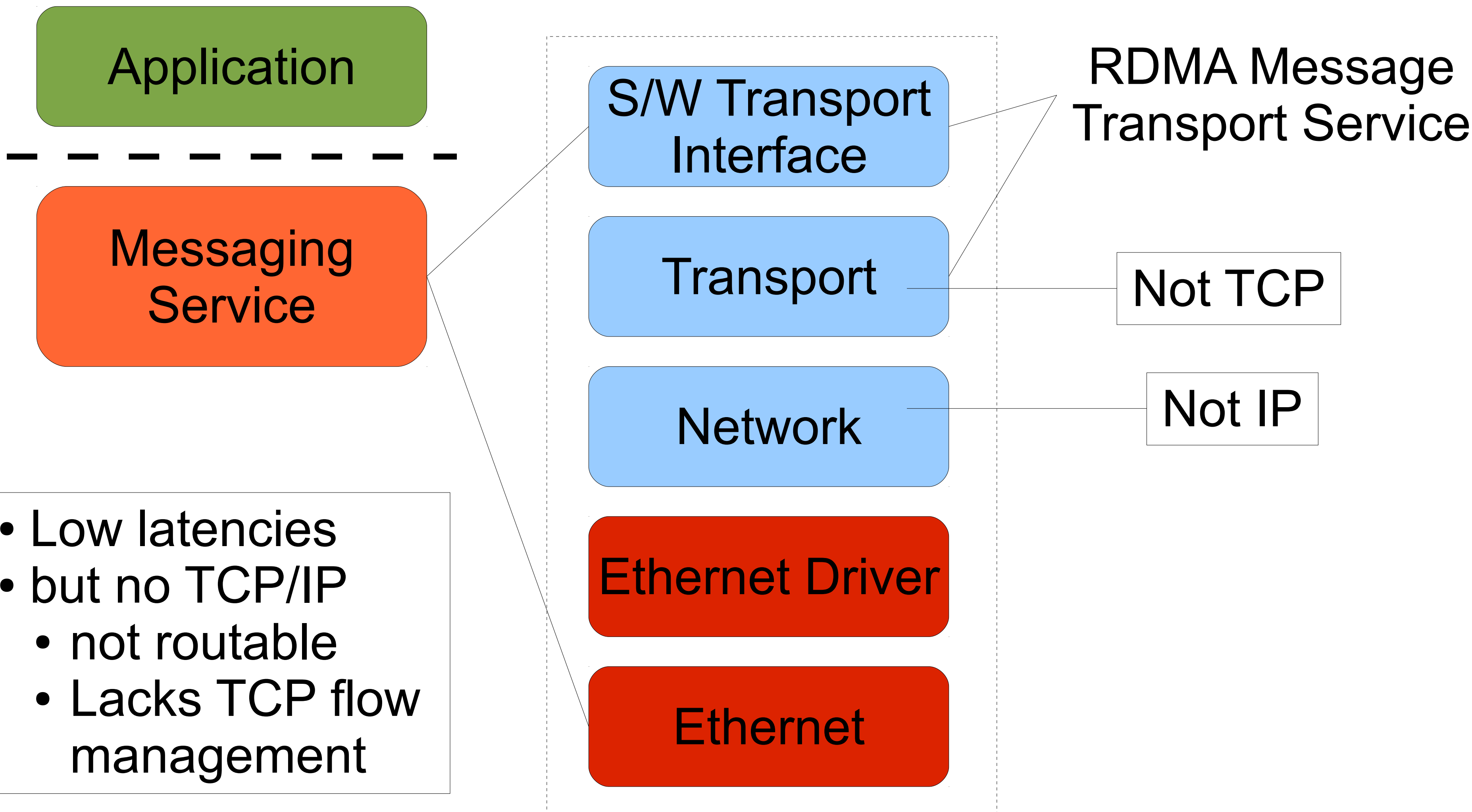
Introduction to Infiniband for End Users: Paul Grun, Infiniband Trade Association

University of BRISTOL

IB versus Ethernet

- Ethernet is commodity (& so cheaper..?)
- Many practitioners have knowledge of ethernet
- Competition between Ethernet and IB around bandwidth: 40 & 100 Gbit/s are in deployed today in IB (QDR in BCp3=32Gbit/s, FDR=54Gbit/s, EDR in BCp4=100Gbit/s)
  - See: https://en.wikipedia.org/wiki/InfiniBand
- But what to do about the TCP/IP latency?

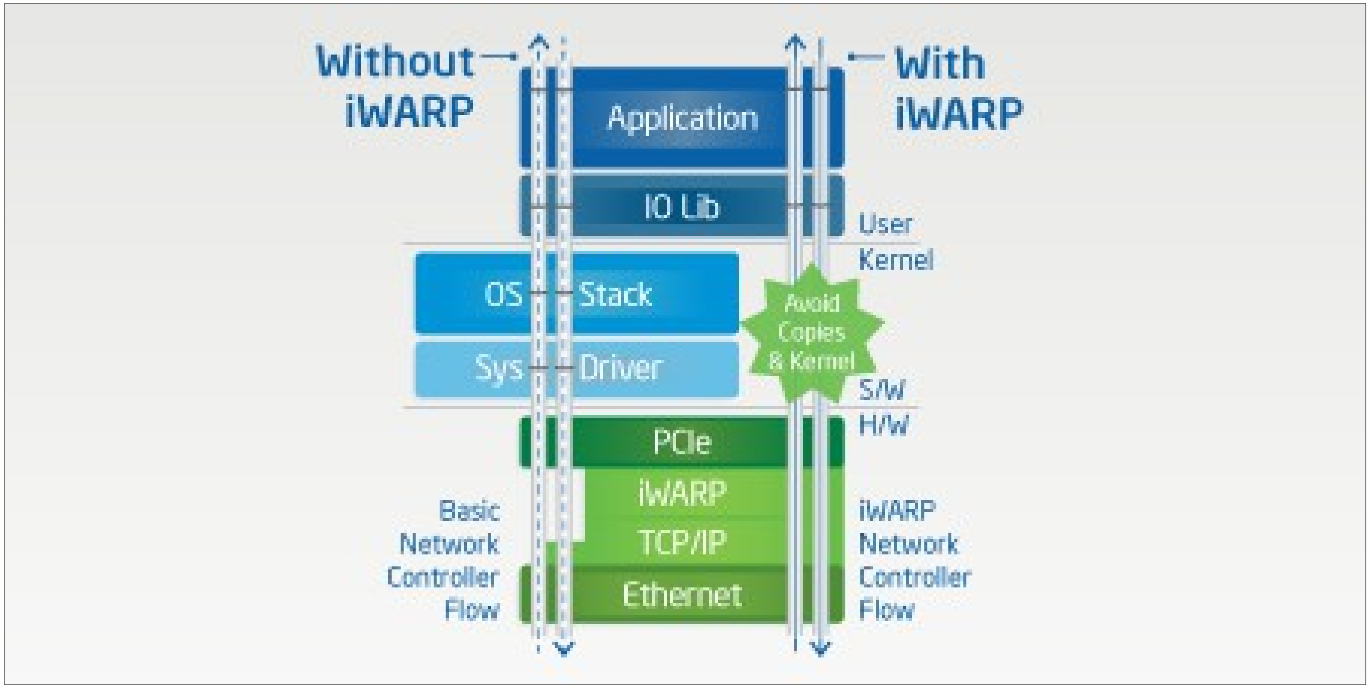
RoCE: RDMA over (Converged) Ethernet



- Low latencies
- but no TCP/IP
  - not routable
  - Lacks TCP flow management

University of BRISTOL

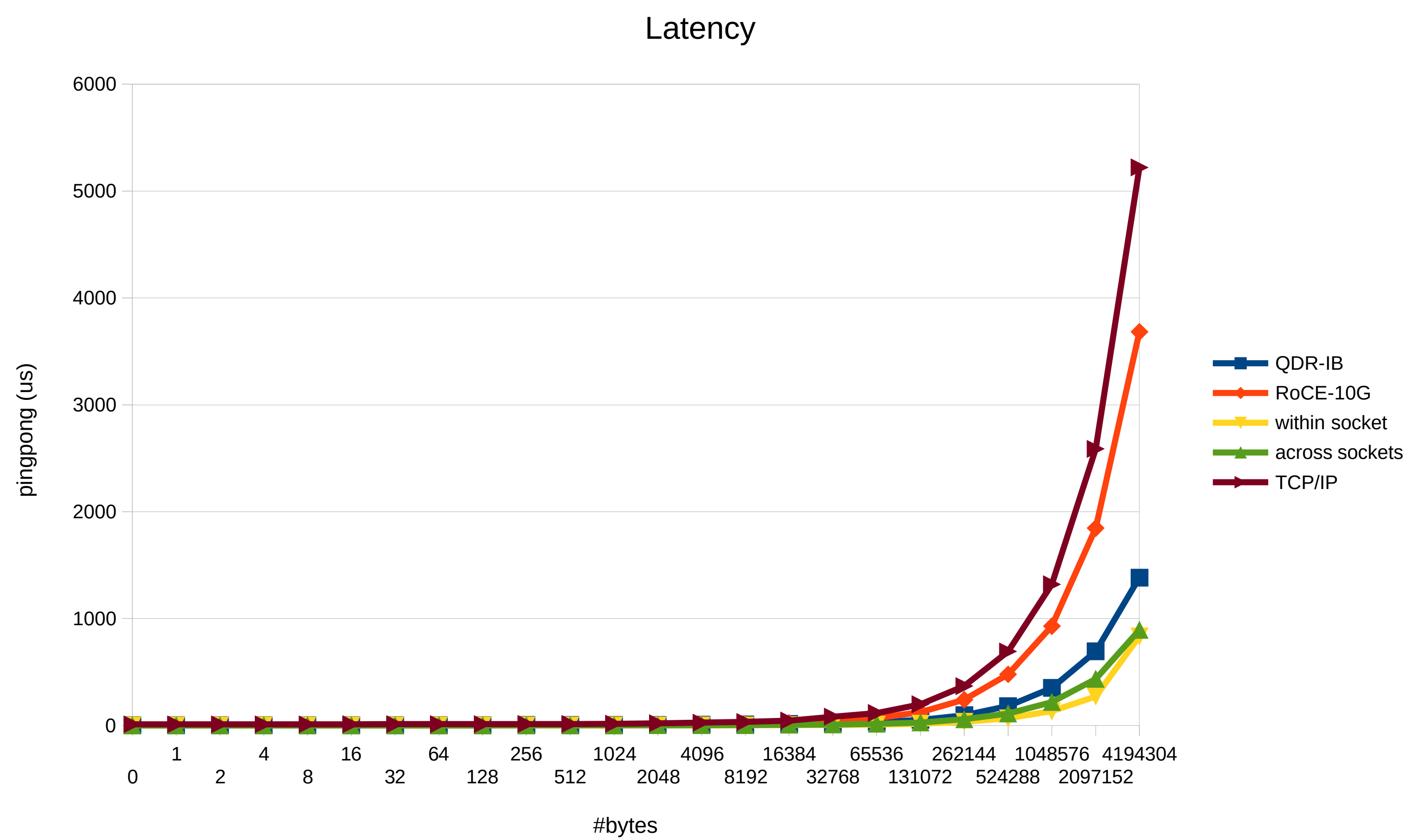
IWARP: Internet Wide Area RDMA Protocol



- TCP/IP is implemented on the NIC
- Routable, but how much extra latency?

University of BRISTOL

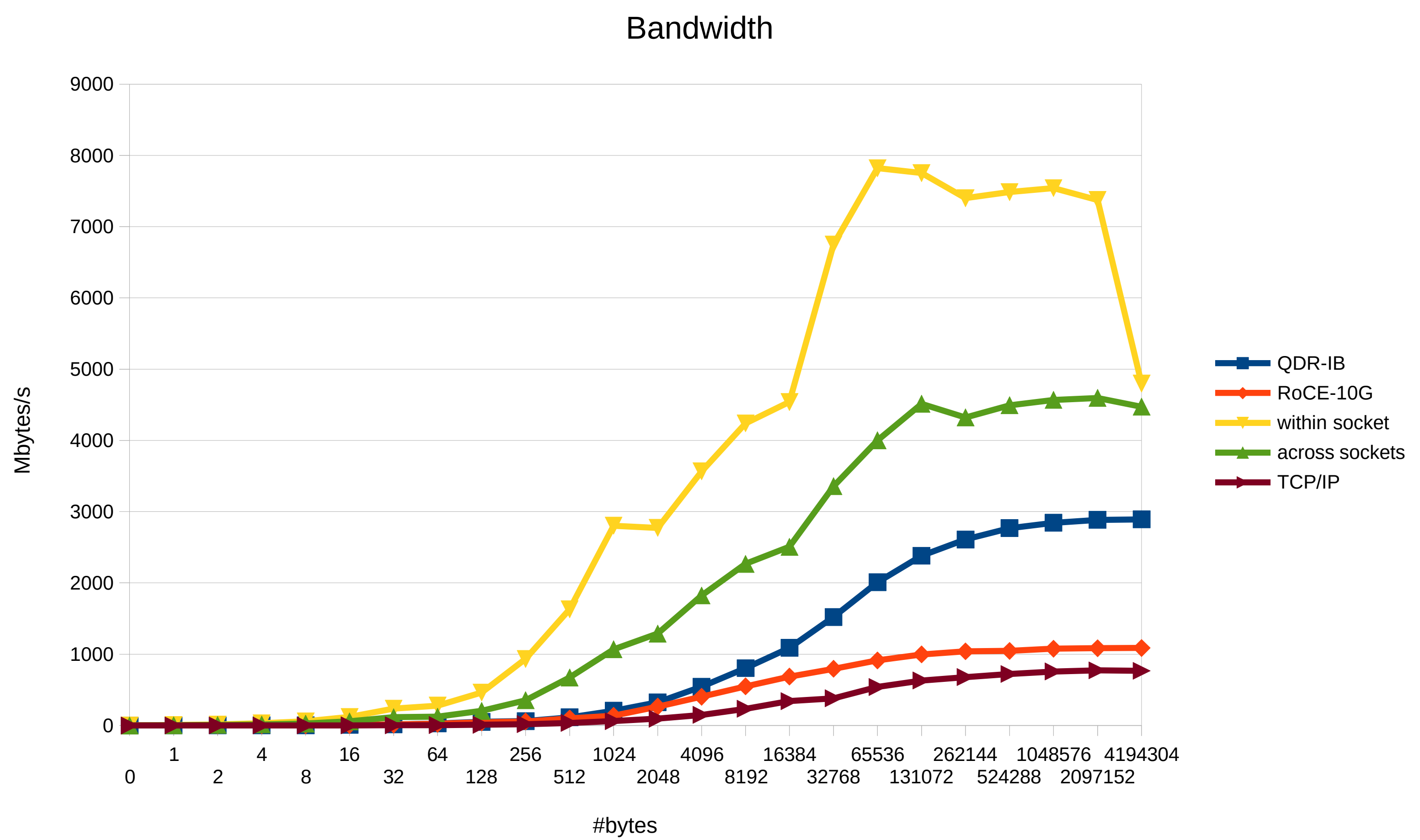
Latency



	within-socket	across-sockets	QDR IB	10G RoCE	tcp/ip
0 byte msg	0.12us	0.25us	2.03us	2.2us	10.5us

QDR from BCp3, RoCE from Emerald

Bandwidth

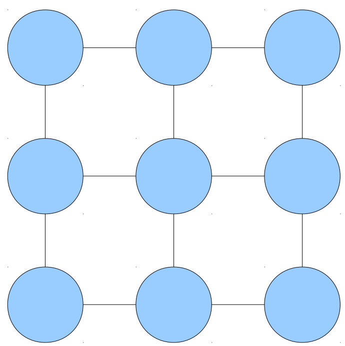


10Gbit/s = 1250Mbytes/s
32Gbit/s = 4000Mbytes/s (QDR IB)

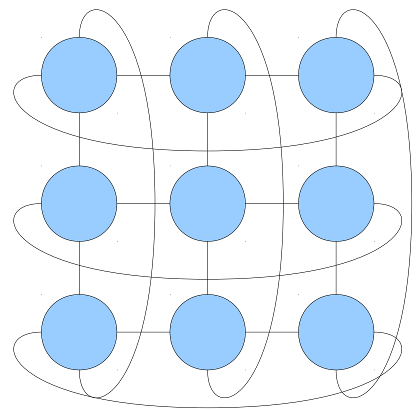
QDR from BCp3, RoCE from Emerald

University of BRISTOL

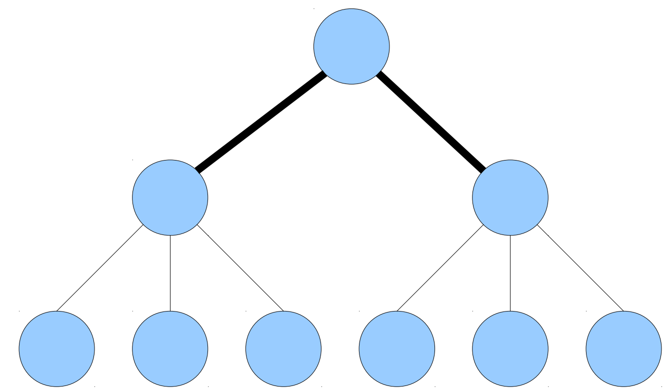
Network Topologies



2d-grid



2d-torus



(fat) tree

Torus – good for neighbour exchanges (e.g. halo-exchange)
Tree – cost effective, good for collective operations

Networking Future

- Bandwidth Improvements:
  - 40 & 100 Gbit/s Ethernet
  - EDR (100Gbit/s), HDR (200Gbit/s) IB
- Latency reductions:
  - Many vendors will be integrating the network adapter into the processor's silicon die
  - This will bring higher performance and also consume power
  - Taking longer than expected though.

University of BRISTOL

Storage

"If you solve a compute problem, you are rewarded with a storage problem!"

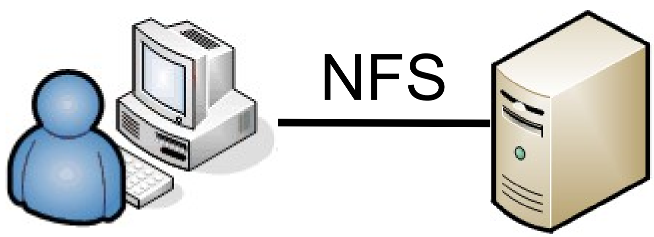
University of BRISTOL

Bottlenecks

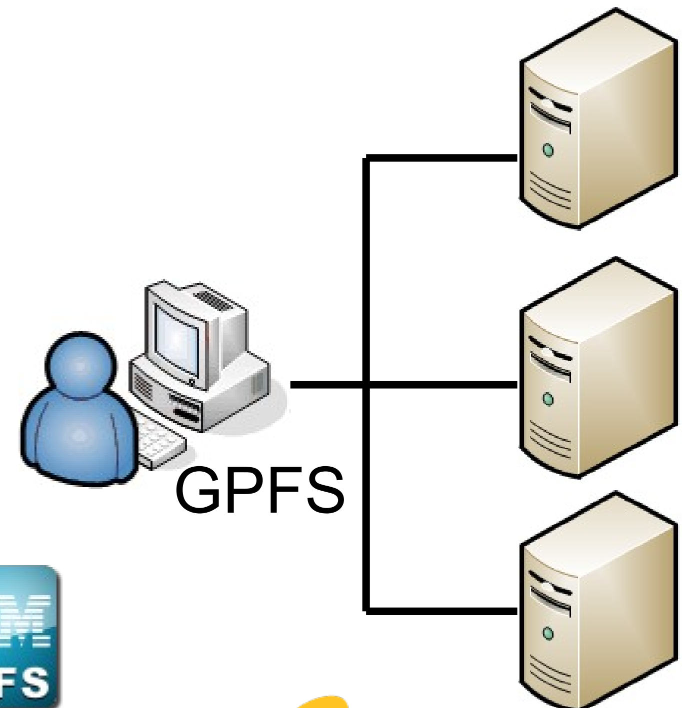
You can have the most processors, fastest network and best code, but your applications will stall creating output or reading input if you don't have commensurate storage

Serial vs. Parallel File Systems

Serial File System



Parallel File System



- Parallel: network/disk load is spread→ faster
- But:
  - Small files are hard to split and so reside on a single disk → back to the serial situation
  - Try not to use small files..

IBM GPFS

panasas

lustre

pNFS

University of BRISTOL

Parallel File System vs Local Disk

watch out for caching effects..

Test1 – 'large' files:
wget https://www.kernel.org/pub/linux/kernel/v3.x/testing/linux-3.10-rc7.tar.xz
time cp linux-3.10-rc7.tar.xz linux-3.10-rc7.tar.xz.copy
1.5s (~50MB/s) vs. 0.7s (~100MB/s)

Test2 – small files:
tar --use-compress-program=xz -xf linux-3.10-rc7.tar.xz (~570MB unpacked)
time cp -r linux-3.10-rc7 linux-3.10-rc7.copy
3m46s (2.5MB/s) vs 4.3s (~133MB/s)

University of BRISTOL

Parallel File System vs Local Disk

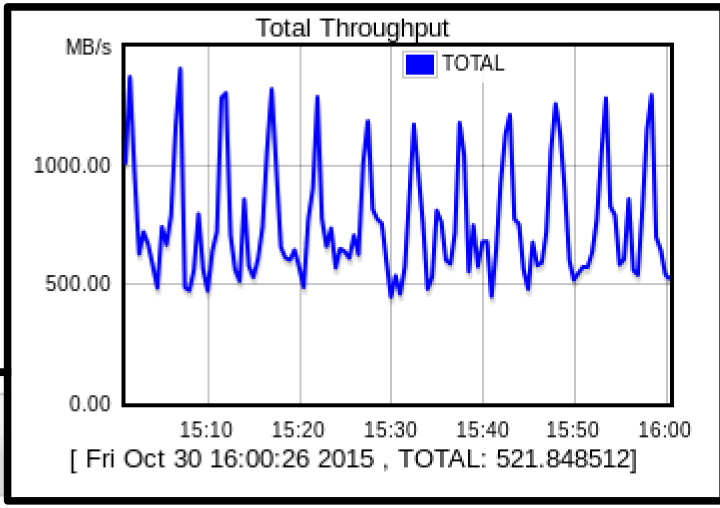
£250,000!

£300?

Things are not looking too good for the parallel file system!

University of BRISTOL

Empirical Disk Comparisons

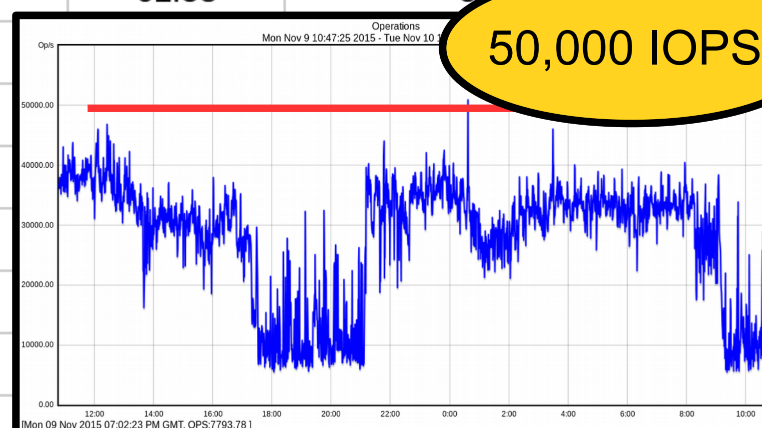


Local disk, using bonnie++:
~120MB/s sequential read
~150 MB/s sequential write
~150 IOPs (random seeks)

Caching can help here

		Ops/sec	MB/sec	Response Time (msec)
10.139.171.114	*newblue1.ib3.cluster*	648	76.76	7.32
10.139.171.117	*newblue4.ib3.cluster*	221	28.41	4.16
10.139.171.115	*newblue2.ib3.cluster*	219	61.58	5.5[illegible]
10.139.1.61	*node45-013.ib3.cluster*	203		
10.139.0.71	*node34-023.ib3.cluster*	201		
10.139.2.12	*node43-012.ib3.cluster*	195		
10.139.0.182	*node32-038.ib3.cluster*	143		
10.139.0.183	*node32-039.i*	114		
10.139.0.43	*node35-04*	113		
10.139.1.3	*node47-003.i*	104		

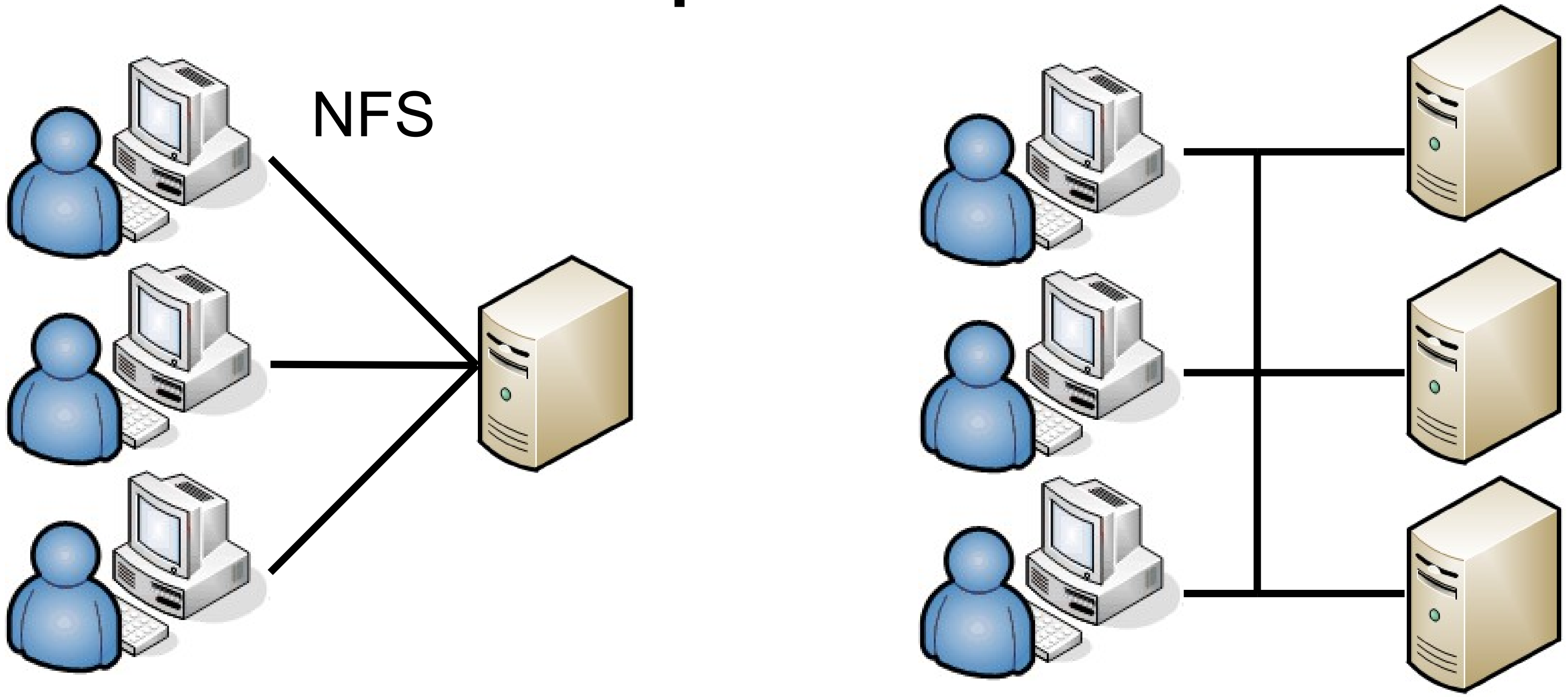
Maybe not so Bad after all..



50,000 IOPS!

University of BRISTOL

Multiple clients



File copying bandwidths for a large file (~350Mb)

	Clients x1	Clients x4	Clients x8
Desktop	~65MB/s	~10MB/s	~0.5MB/s
Parallel fs	~160MB/s	~130MB/s	~67MB/s

University of BRISTOL

Parallel File System vs Local Disk

- So where does that leave us?
  - We *need* a shared filesystem on a cluster (think home directories) and parallel file systems provide (very) good aggregate performance.
  - Parallel file systems can deliver bandwidths far in excess of that of a single disk (or SSD).
  - Since the shared file-system is at the far end of a network connection, latency has to suffer. For *some* workloads this implies that would do well to use of both kinds of storage on offer.

University of BRISTOL

HPC Filesystem – A Quiz

- ~85TB of capacity on BCp2
  - How many files in total?
  - What % of files 0-2KB?
  - What % of files <= 64KB?

HPC Filesystem – A Quiz

- ~85TB of capacity on BCp2
  - How many files in total? **60Million**
  - What % of files 0-2KB?
  - What % of files <= 64KB?

University of BRISTOL

HPC Filesystem – A Quiz

- ~85TB of capacity on BCp2
  - How many files in total? **60Million**
  - What % of files 0-2KB? **33%!**
  - What % of files <= 64KB?

University of BRISTOL

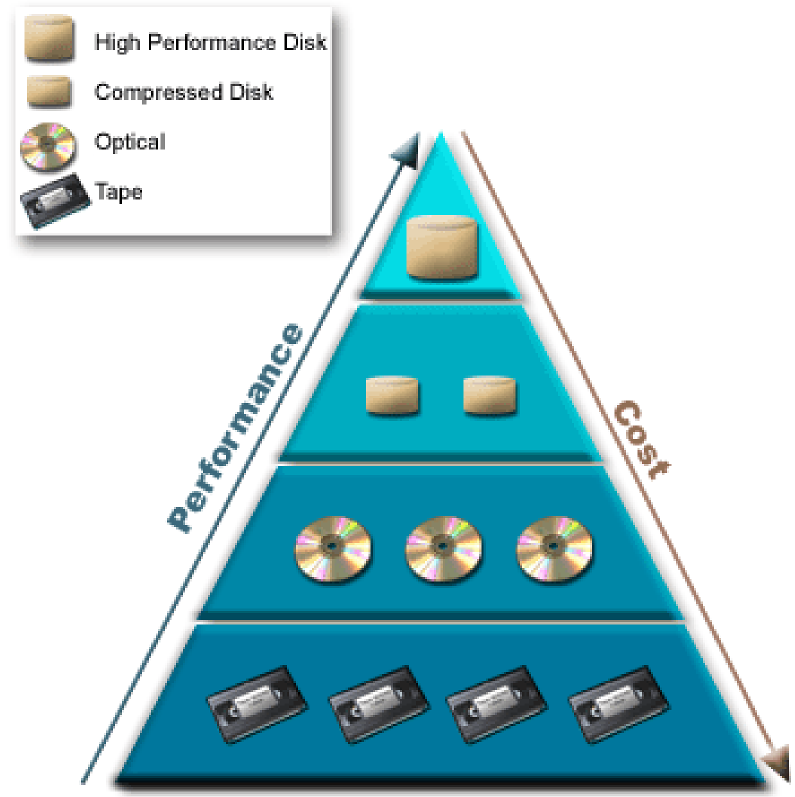
HPC Filesystem – A Quiz

- ~85TB of capacity on BCp2
  - How many files in total? **60Million**
  - What % of files 0-2KB? **33%!**
  - What % of files <= 64KB? **~76%**
- 64KB is the stripe size for, e.g., Panasas, i.e. no parallel gain for ¾ of the files on the system.
- This is not atypical for an HPC filesystem.
- You're the next generation—spread the word!

University of BRISTOL

Hierarchical Storage Management (HSM)

- Applications read from, and write to fastest storage medium
- Policies are used to automatically demote cohorts of 'cold' data and promote cohorts of 'hot' data.



University of BRISTOL

Summary

- We need to see systems in the round:
  - Networking and storage are key to HPC, just as are processors and access to memory.
- HPC networks are designed for low latency & high bandwidth, but will always be limited on both of these dimensions.
- Parallel file systems can offer better performance—if used intelligently.
- HSM is driven by a desire for performance but also storage economics.

University of BRISTOL