HPC Parallel Programming: Introduction to Multinode Programming

Parallelization and Optimization Group
TATA Consultancy Services, SahyadriPark Pune, India

May 2, 2013

HPC Computing Cluster:

Figure: High Performance Multicore Multinode Cluster:

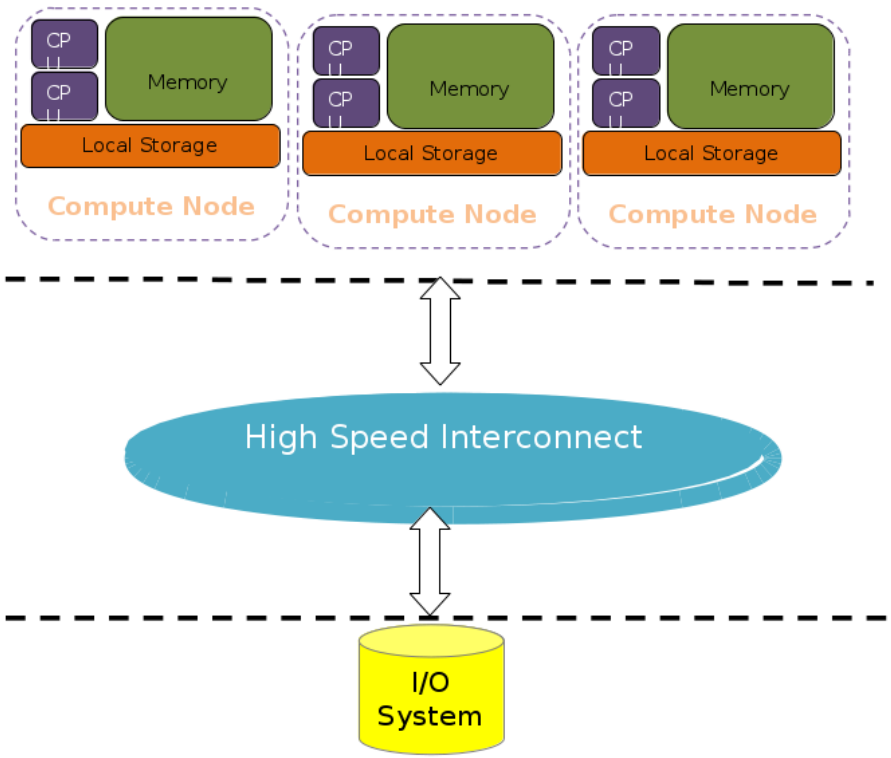
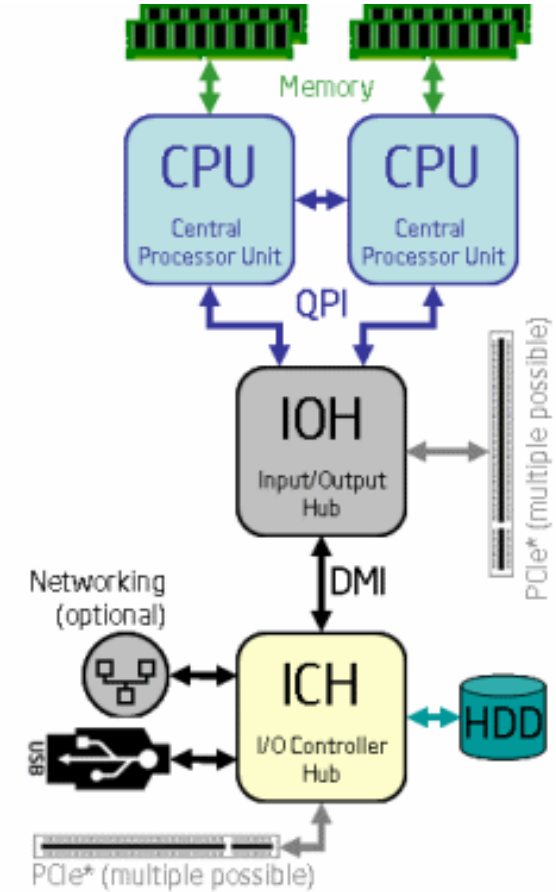


Figure: Courtesy Sanket Sinha from Data Operations

Memory Access:

Figure: CPU to Memory connection
NUMA Source: www.intel.com



Memory Access:

Figure: CPU to Memory connection NUMA Source: www.intel.com

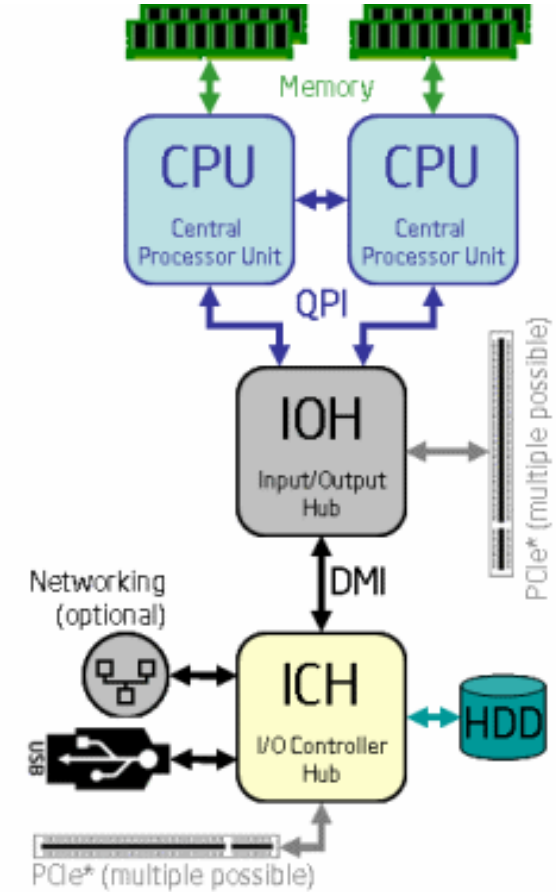
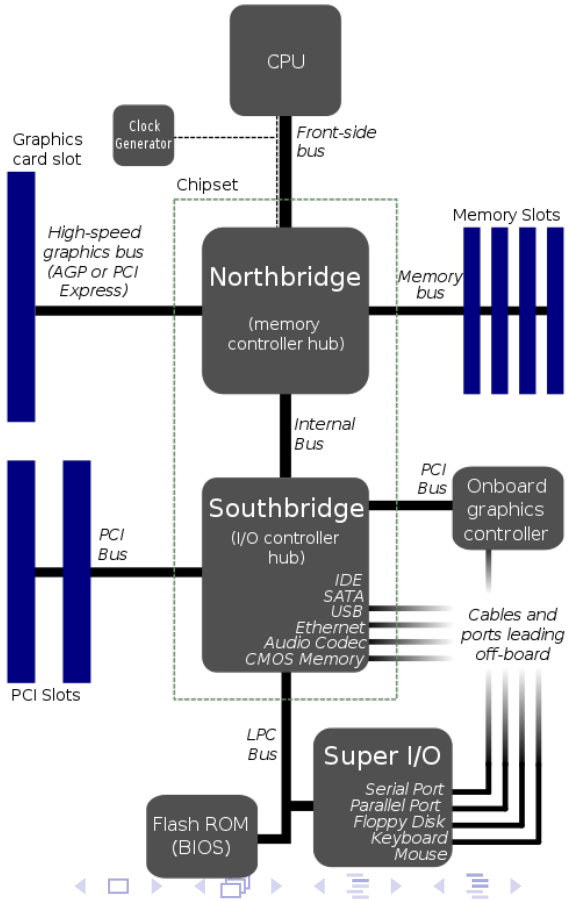


Figure: CPU to Memory connection via FrontSide Bus. Source: Wikipedia



CPU Memory Bandwidth: Sandy Bridge ES 2670

CPU Specs		Comment
No of Sockets	2	
Technology	32 nm	
No. of Cores	8	
Clock Rate	2.6 Ghz	
No. of Floating Point operations per clock	8	8*3*8 = 192 2.6 * 192= 499.2

CPU Memory Bandwidth: Sandy Bridge ES 2670

CPU Specs		Comment
No of Sockets	2	
Technology	32 nm	
No. of Cores	8	
Clock Rate	2.6 Ghz	
No. of Floating Point operations per clock per core	8	8*3*8 = 192 2.6 * 192= 499.2 499.2 * 8 = 3993.6 GB/s

CPU Memory Bandwidth: Sandy Bridge ES 2670

CPU Specs		Comment
No of Sockets	2	
Technology	32 nm	
No. of Cores	8	
Clock Rate	2.6 Ghz	
No. of Floating Point operations per clock per core	8	8*3*8 = 192 2.6 * 192= 499.2 499.2 * 8 = 3993.6 GB/s
QPI speed	8GT/s	
PCI Express 3	40 lane	

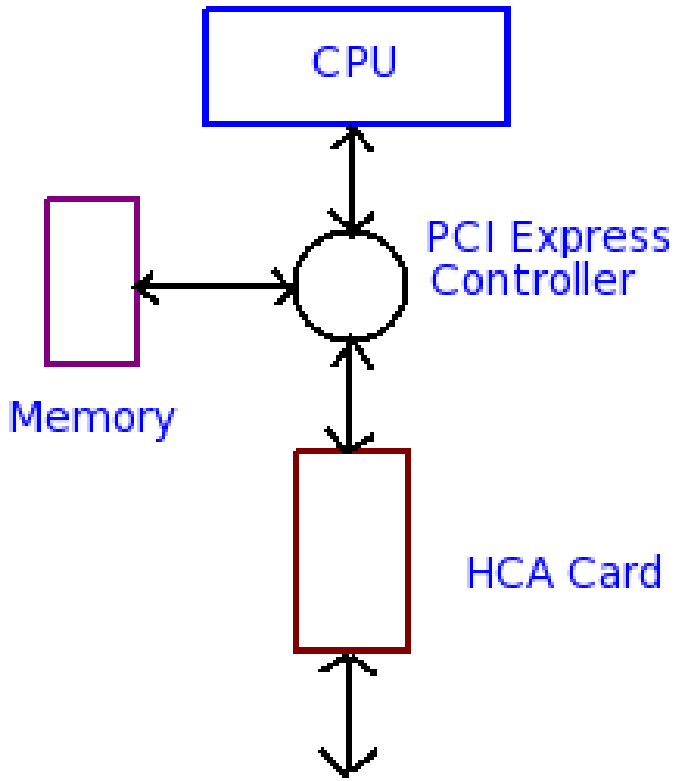
CPU Memory Bandwidth: Sandy Bridge ES 2670

CPU Specs		Comment
No of Sockets	2	
Technology	32 nm	
No. of Cores	8	
Clock Rate	2.6 Ghz	
No. of Floating Point operations per clock per core	8	8*3*8 = 192 2.6 * 192= 499.2 499.2 * 8 = 3993.6 GB/s
QPI speed	8GT/s	
PCI Express 3	40 lane	

Mem Specs		Comment
Memory Type	DDR3-800/ 1066/1333/ 1600	1333 * 8 bytes
No. of Channels	4	allows for parallel reads by the cpu
Memory CPU bus width	64 bits	
Max Memory Bandwidth	51.2GB/s	1333 * 8 * 4 = 42.656 GB/s
Max Memory Size	750 GB	

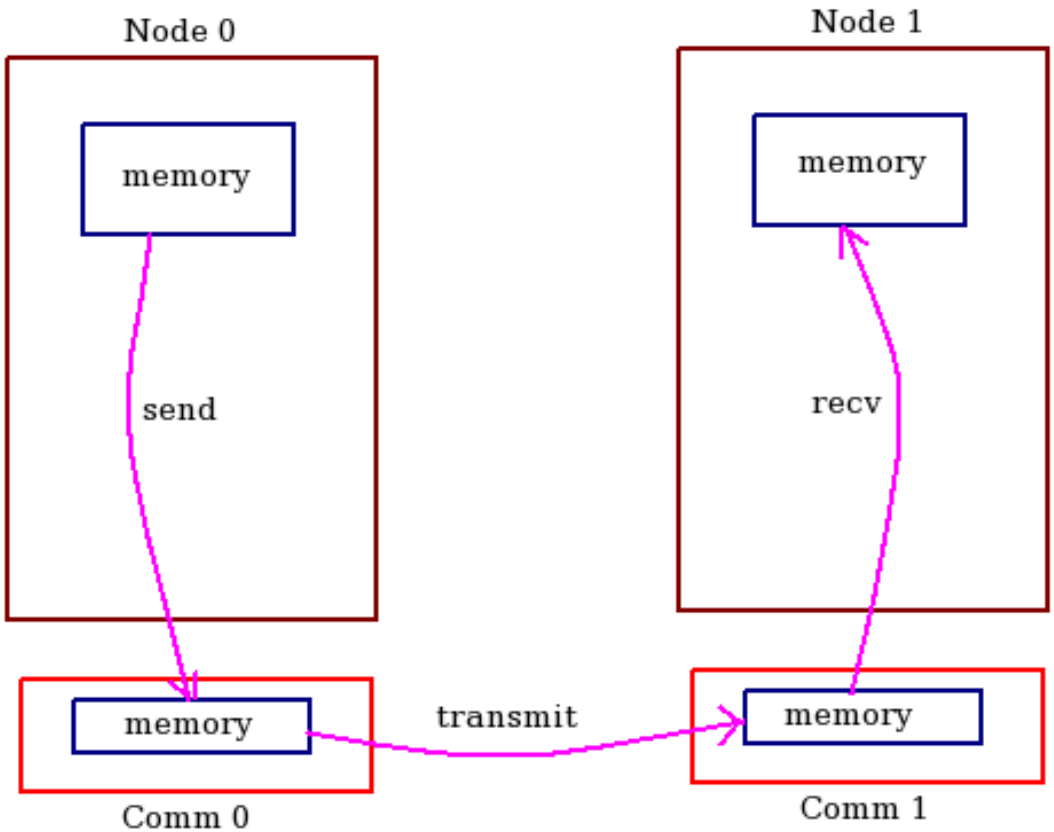
There is 100X gap between the CPU and Memory Bandwidth.

Communication Architecture:



Communication Architecture:HCA port

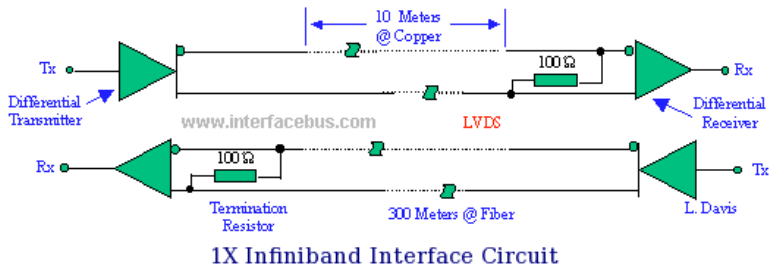
MPI Send/ Receive



1. Total communication time: send+transmit+receive
2. Speed up possible by splitting into packets?
3. Other user methods to speed up

Infiniband Description

Figure: Infiniband physical layer:



1X Infiniband Interface Circuit

Figure: Infiniband Protocol:

Infiniband packet structure

1. Point to Point Protocol
2. Bidirectional links per channel, differential pair per direction
3. Messages are converted to packets
4. Packet has various fields
5. Routing
6. Error correction

Source: www.interface-bus.com

Infiniband Description:

Table: InfiniBand Signaling Rate

links	SDR	DDR	QDR	FDR	EDR
1x	2.5Gbps	5Gbps	10.0Gbps	14.0625	28.78125Gbps
4x	10Gbps	20Gbps	40Gbps	56.25	115.125Gbps
12x	30Gbps	60Gbps	120Gbps	168.75	345.375Gbps

Infiniband Description:

Table: InfiniBand Signaling Rate

links	SDR	DDR	QDR	FDR	EDR
1x	2.5Gbps	5Gbps	10.0Gbps	14.0625	28.78125Gbps
4x	10Gbps	20Gbps	40Gbps	56.25	115.125Gbps
12x	30Gbps	60Gbps	120Gbps	168.75	345.375Gbps

Source: Wikipedia

1. Every 10 bits have 2 error correcting bits in SDR, DDR and QDR.
2. Additional control bits are there in the overall frame.

Infiniband Description:

Table: InfiniBand Signaling Rate

links	SDR	DDR	QDR	FDR	EDR
1x	2.5Gbps	5Gbps	10.0Gbps	14.0625	28.78125Gbps
4x	10Gbps	20Gbps	40Gbps	56.25	115.125Gbps
12x	30Gbps	60Gbps	120Gbps	168.75	345.375Gbps

1. Every 10 bits have 2 error correcting bits in SDR, DDR and QDR.
2. Additional control bits are there in the overall frame.

Source: Wikipedia

Table: Infiniband Latency:

SDR	200 ns.
DDR	140 ns.
QDR	100 ns.

Table: MPI Latency:

Mellanox	1.29 micro secs.
Qlogic	2.6 micro secs.

Infiniband Description:

Table: InfiniBand Signaling Rate

links	SDR	DDR	QDR	FDR	EDR
1x	2.5Gbps	5Gbps	10.0Gbps	14.0625	28.78125Gbps
4x	10Gbps	20Gbps	40Gbps	56.25	115.125Gbps
12x	30Gbps	60Gbps	120Gbps	168.75	345.375Gbps

1. Every 10 bits have 2 error correcting bits in SDR, DDR and QDR.
2. Additional control bits are there in the overall frame.

Source: Wikipedia

Table: Infiniband Latency:

SDR	200 ns.
DDR	140 ns.
QDR	100 ns.

Table: MPI Latency:

Mellanox	1.29 micro secs.
Qlogic	2.6 micro secs.

Fundamental Challenge:There is 10X bandwdith gap between the cpu memory bandwidth and across board bandwdith. There is over 10 X gap between latency between memory to cpu .1 μ secs and across board latency of 1.29 μsecs.

Thank You!