

Network Performance Benchmarks

Mellanox 40GE NICs, Caltech Fast Data Transfer (FDT) Application and MonALISA

NOV 2010

**Azher Mughal
Caltech**

Servers and Network Test Setup

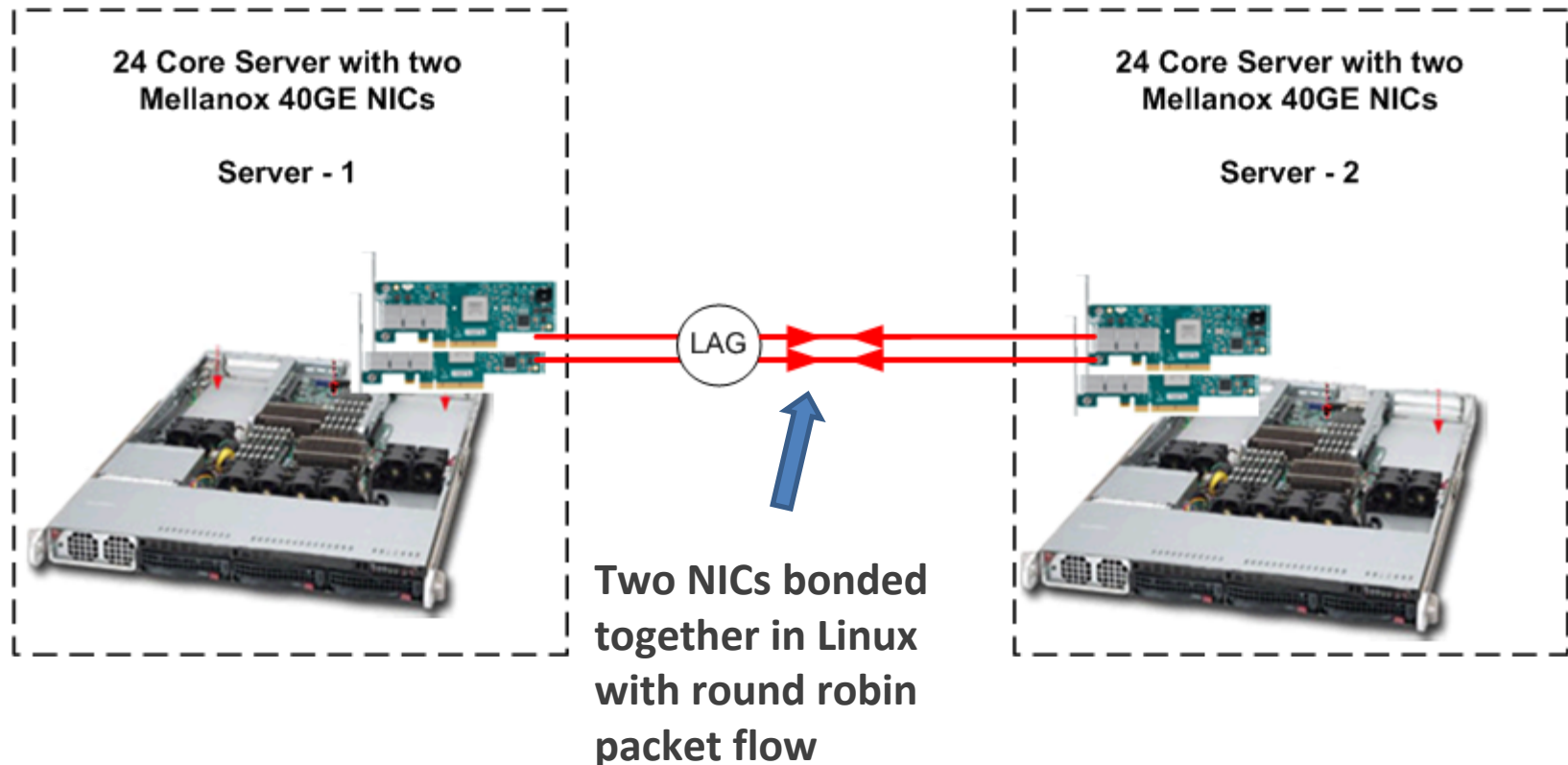


Server:

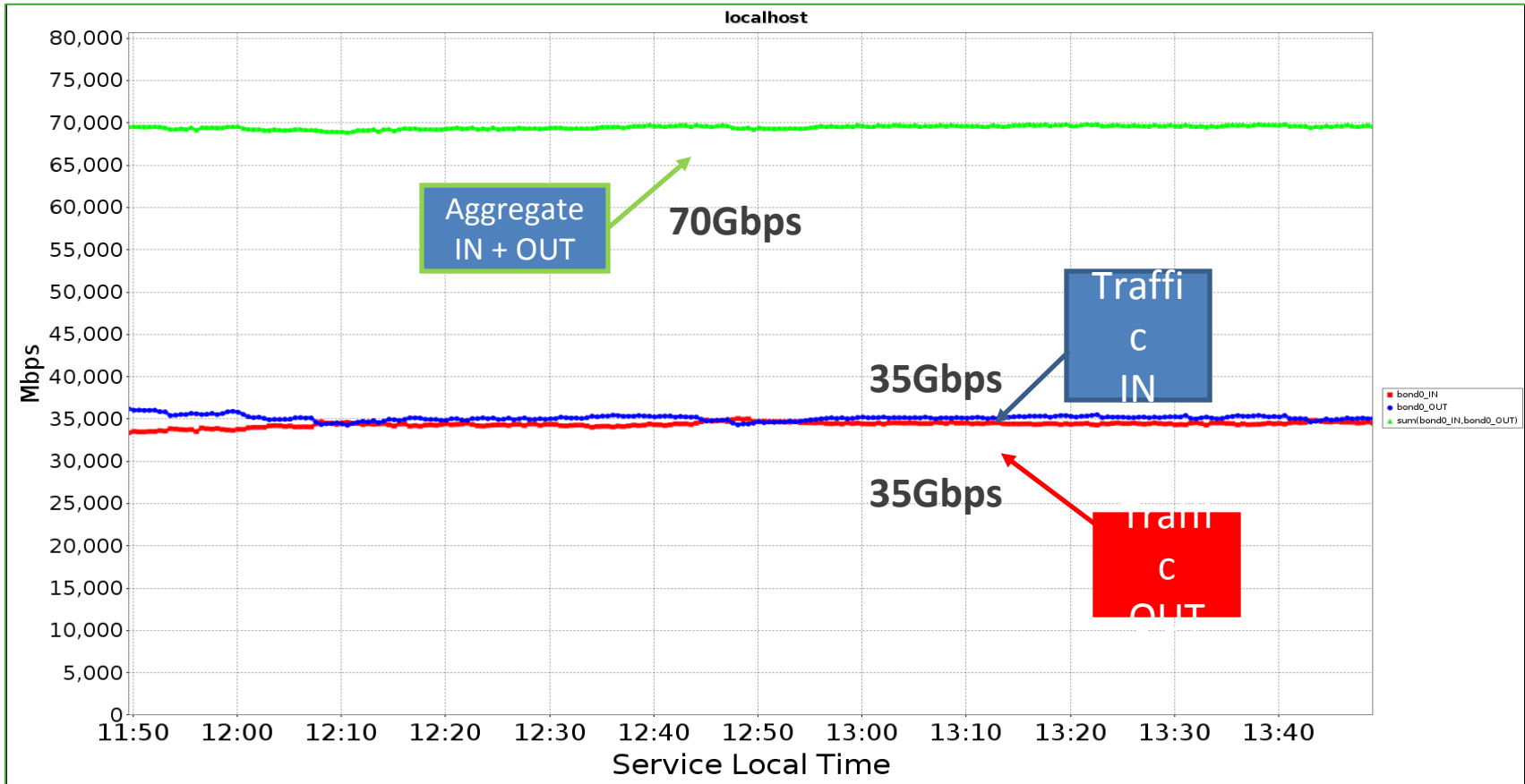
SuperMicro 6016XT-TF
Dual Intel X5670, 24GB DDR3 RAM
Supporting 4 Gen2.0 x8 slots
OS: RedHat Linux, Kernel 2.6.36

Mellanox Driver:

16 Receive queues + 16 Transmit queues
Challenge: Optimize queues to CPU cores

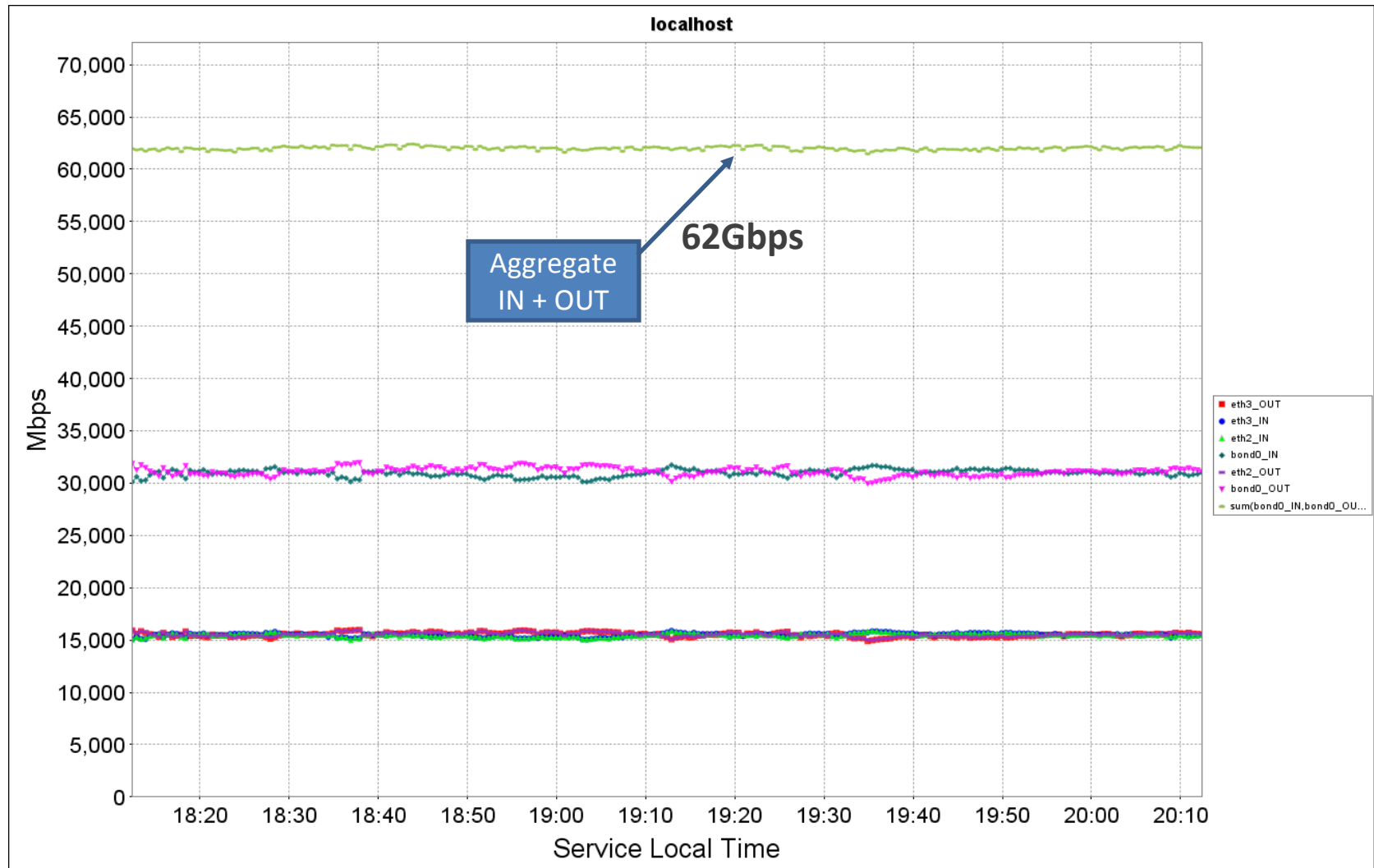


Network Test: Two 40GE bonded NIC



FDT "nettest" buffer tests, 16 streams in each direction

Memory Test: Two 40GE bonded NIC



FDT reading from /dev/zero and writing to /dev/null memory tests, 16 streams in each direction

UltraLight Kernel 2.6.36-UL3

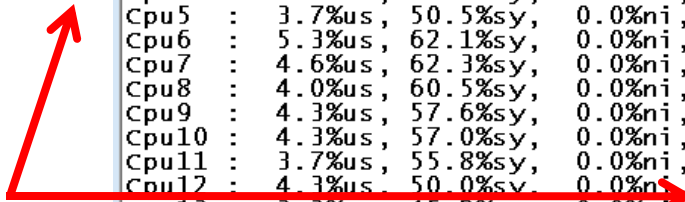


```
top - 17:29:28 up 12:15, 4 users, load average: 9.15, 8.52, 8.42
Tasks: 188 total, 8 running, 180 sleeping, 0 stopped, 0 zombie
Cpu0 : 4.3%us, 53.8%sy, 0.0%ni, 36.8%id 0.0%wa, 0.0%hi, 5.0%si, 0.0%st
Cpu1 : 1.0%us, 11.9%sy, 0.0%ni, 14.6%id 0.0%wa, 0.0%hi, 72.5%si, 0.0%st
Cpu2 : 0.3%us, 5.0%sy, 0.0%ni, 16.9%id 0.0%wa, 0.0%hi, 77.7%si, 0.0%st
Cpu3 : 2.3%us, 20.3%sy, 0.0%ni, 28.9%id 0.0%wa, 0.0%hi, 48.5%si, 0.0%st
Cpu4 : 1.3%us, 12.5%sy, 0.0%ni, 20.8%id 0.0%wa, 0.0%hi, 65.3%si, 0.0%st
Cpu5 : 4.0%us, 43.2%sy, 0.0%ni, 49.2%id 0.0%wa, 0.0%hi, 3.7%si, 0.0%st
Cpu6 : 4.7%us, 64.3%sy, 0.0%ni, 23.0%id 0.0%wa, 0.0%hi, 8.0%si, 0.0%st
Cpu7 : 5.0%us, 61.9%sy, 0.0%ni, 24.2%id 0.0%wa, 0.0%hi, 8.9%si, 0.0%st
Cpu8 : 4.0%us, 60.3%sy, 0.0%ni, 27.7%id 0.0%wa, 0.0%hi, 8.0%si, 0.0%st
Cpu9 : 4.0%us, 60.3%sy, 0.0%ni, 32.0%id 0.0%wa, 0.0%hi, 6.3%si, 0.0%st
Cpu10 : 4.0%us, 60.3%sy, 0.0%ni, 36.0%id 0.0%wa, 0.0%hi, 6.7%si, 0.0%st
Cpu11 : 4.0%us, 60.3%sy, 0.0%ni, 38.9%id 0.0%wa, 0.0%hi, 6.3%si, 0.0%st
Cpu12 : 4.0%us, 60.3%sy, 0.0%ni, 48.3%id 0.0%wa, 0.0%hi, 4.0%si, 0.0%st
Cpu13 : 2.3%us, 42.3%sy, 0.0%ni, 51.3%id 0.0%wa, 0.0%hi, 4.0%si, 0.0%st
Cpu14 : 3.3%us, 39.1%sy, 0.0%ni, 54.3%id 0.0%wa, 0.0%hi, 3.3%si, 0.0%st
Cpu15 : 3.3%us, 37.5%sy, 0.0%ni, 56.1%id 0.0%wa, 0.0%hi, 3.0%si, 0.0%st
Cpu16 : 2.0%us, 33.6%sy, 0.0%ni, 62.1%id 0.0%wa, 0.0%hi, 2.3%si, 0.0%st
Cpu17 : 2.3%us, 35.8%sy, 0.0%ni, 58.3%id 0.0%wa, 0.0%hi, 3.6%si, 0.0%st
Cpu18 : 3.7%us, 52.0%sy, 0.0%ni, 38.3%id 0.0%wa, 0.0%hi, 6.0%si, 0.0%st
Cpu19 : 3.3%us, 49.5%sy, 0.0%ni, 40.5%id 0.0%wa, 0.0%hi, 6.7%si, 0.0%st
Cpu20 : 2.7%us, 46.8%sy, 0.0%ni, 44.2%id 0.0%wa, 0.0%hi, 6.3%si, 0.0%st
Cpu21 : 2.7%us, 43.3%sy, 0.0%ni, 47.3%id 0.0%wa, 0.0%hi, 6.7%si, 0.0%st
Cpu22 : 2.7%us, 39.7%sy, 0.0%ni, 52.0%id 0.0%wa, 0.0%hi, 5.7%si, 0.0%st
Cpu23 : 2.7%us, 39.0%sy, 0.0%ni, 52.7%id 0.0%wa, 0.0%hi, 5.7%si, 0.0%st
Mem: 24730968k total, 2113320k used, 22617648k free, 61656k buffers
Swap: 16779888k total, 0k used, 16779888k free, 378572k cached
```

Server - 1

Requires manual IRQ shift across CPU Cores

Improved CPU idle time and load balancing across cores



```
top - 17:29:28 up 12:15, 4 users, load average: 9.83, 10.51, 10.61
Tasks: 188 total, 8 running, 180 sleeping, 0 stopped, 0 zombie
Cpu0 : 24.5%id, 0.0%wa, 0.0%hi, 28.5%si, 0.0%st
Cpu1 : 12.9%id, 0.0%wa, 0.0%hi, 68.5%si, 0.0%st
Cpu2 : 10.6%id, 0.0%wa, 0.3%hi, 77.7%si, 0.0%st
Cpu3 : 12.3%id, 0.0%wa, 0.0%hi, 86.3%si, 0.0%st
Cpu4 : 16.3%id, 0.0%wa, 0.0%hi, 79.7%si, 0.0%st
Cpu5 : 41.2%id, 0.0%wa, 0.0%hi, 4.7%si, 0.0%st
Cpu6 : 23.3%id, 0.0%wa, 0.0%hi, 9.3%si, 0.0%st
Cpu7 : 24.5%id, 0.0%wa, 0.0%hi, 7.0%si, 0.0%st
Cpu8 : 25.9%id, 0.0%wa, 0.0%hi, 7.0%si, 0.0%st
Cpu9 : 30.1%id, 0.0%wa, 0.0%hi, 7.0%si, 0.0%st
Cpu10 : 30.3%id, 0.0%wa, 0.0%hi, 7.0%si, 0.0%st
Cpu11 : 33.6%id, 0.0%wa, 0.0%hi, 7.0%si, 0.0%st
Cpu12 : 41.3%id, 0.0%wa, 0.0%hi, 4.3%si, 0.0%st
Cpu13 : 47.2%id, 0.0%wa, 0.0%hi, 4.3%si, 0.0%st
Cpu14 : 49.7%id, 0.0%wa, 0.0%hi, 3.7%si, 0.0%st
Cpu15 : 54.7%id, 0.0%wa, 0.0%hi, 3.3%si, 0.0%st
Cpu16 : 56.7%id, 0.0%wa, 0.0%hi, 3.7%si, 0.0%st
Cpu17 : 50.8%id, 0.0%wa, 0.0%hi, 3.7%si, 0.0%st
Cpu18 : 31.5%id, 0.0%wa, 0.0%hi, 7.9%si, 0.0%st
Cpu19 : 34.9%id, 0.0%wa, 0.0%hi, 8.0%si, 0.0%st
Cpu20 : 35.3%id, 0.0%wa, 0.0%hi, 8.7%si, 0.0%st
Cpu21 : 40.3%id, 0.0%wa, 0.0%hi, 7.3%si, 0.0%st
Cpu22 : 40.3%id, 0.0%wa, 0.0%hi, 7.0%si, 0.0%st
Cpu23 : 46.0%id, 0.0%wa, 0.0%hi, 6.7%si, 0.0%st
Mem: 24730968k total, 2061352k used, 22669616k free, 58712k buffers
Swap: 0k total, 0k used, 0k free, 368608k cached
```

Server - 2

Summary



- FDT Network Tests utilizes the available PCI-Express Gen 2.0 x8 bandwidth available to **1 x 40GE NIC** to a maximum of **46Gbps** (IN+OUT).
- FDT Network Tests with **2 x 40GE NICs** bonded together achieves a maximum of **70Gbps** (IN+OUT).
- FDT Memory Tests achieves a maximum of **62Gbps** (IN +OUT) when reading from memory based source /dev/zero and dropping packets to /dev/null.
- **Note:** **PCI-e Gen 2.0 x8** slot offers ~24.6 Gbps real world throughput in one direction with about 7Gbps overhead due to 8b/10b encoding (20% loss per direction) and PCI-E layered architecture.
- **Note:** These transfer rates were achieved after careful IRQ steering among the CPU Cores (Dual Xeon X5670, 24 total cores).
- There is still room for more improvements in terms of Kernel, IRQ assignment and FDT.
- PCI-E 3.0 expected early 2011 will use 128b/130b encoding thus reducing loss and providing an effective double rate than PCI-E 2.0 at 1GB/s per lane or 64Gbps FD in an x8 slot.