G.F.Fortugno

Piano per HPC - LNF

Costruzione di un nuovo centro di calcolo basato su tecnologia HPC

Composizione della squadra in termini di istituzioni	**Tecnici e tecnologi dell INFN**
Istituzione Capofila	**INFN**
Budget Totale	**Non ancora definito. Dipenderà dagli altri enti coinvolti**
Budget per i laboratori	**4 Milioni di Euro richiesti al PNRR**
Squadra dei laboratori	**Servizio Tecnico - Computing KLOE - Computing GRID**
Cose da fare HW nei laboratori	**Portare la sala KLOE A250 kW di potenza** **Realizzare uno spazio comune per i CED che si riunificheranno in uno solo CED dedicato al calcolo scientifico**
Richieste di nuovo personale	**Ogni gruppo richiede unità di personale dedicate** **Per il nuovo centro HPC si richiedono almeno 3 unità di personale sistemistico al momento della erogazione del servizio**
Cose da rendicontare a livello di laboratorio	**Tutte le opere necessaria previste dal Servizio Tecnico e ogni elemento del nuovo CED HPC in costruzione**

Tipo di programma:	Tipo di Architettura elaboratore
Single core program - esecuzione su ogni elemento del cluster o dei cluster	Serial program cluster RJE - GRID
Embarassing Parallel - esecuzione dello stesso codice ripetuto su ogni core del cluster	HPC RJE - GRID
Parallelo a grana grossa - Esecuzione di un programma su ogni core di un singolo elaboratore, ripetuto su ogni elaboratore	**HPC EuPRAXIA - MPI - Seriale con riserva del nodo**
Parallelo a grana fine - Un solo programma che gira monopolizzando l'intero cluster	**HPC Spazio/EuPRAXIA - MPI - Full Parallel**

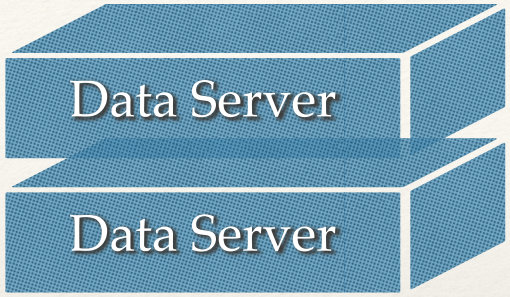
High Performance Computing in LNF

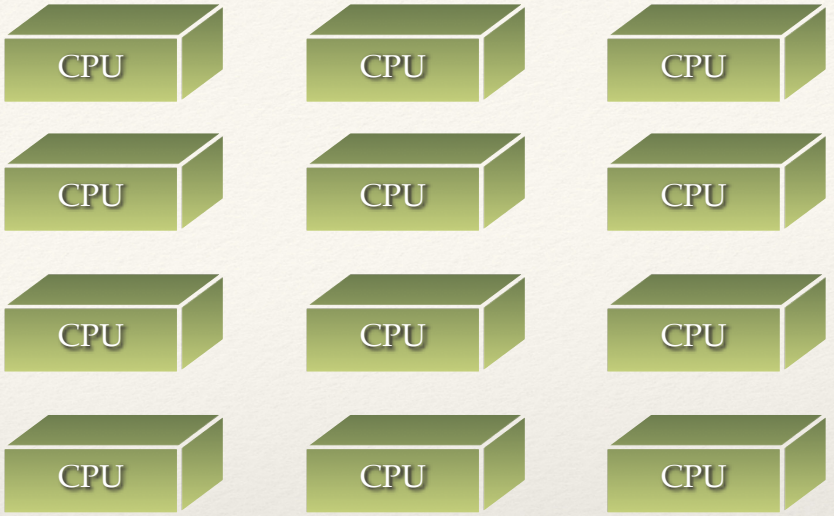
	HPC KLOE	HPC EuPRAXIA	HPC Spazio
Core	1200	**600 + 12 GPU**	25000 + 100 GPU
Tape library	11 Petabyte	**per ora solo backup**	backup e archive 200 Petabyte
Disk Array	1,3 Petabyte	**per ora 300 Terabyte**	100 Petabyte
Network	Ethernet - Fibre Channel	**Ethernet - Fibre Channel - Indiniband**	Ethernet - Fibre Channel - Indiniband

KLOE Experiment

Embarassing Parallel

Dati su NFS



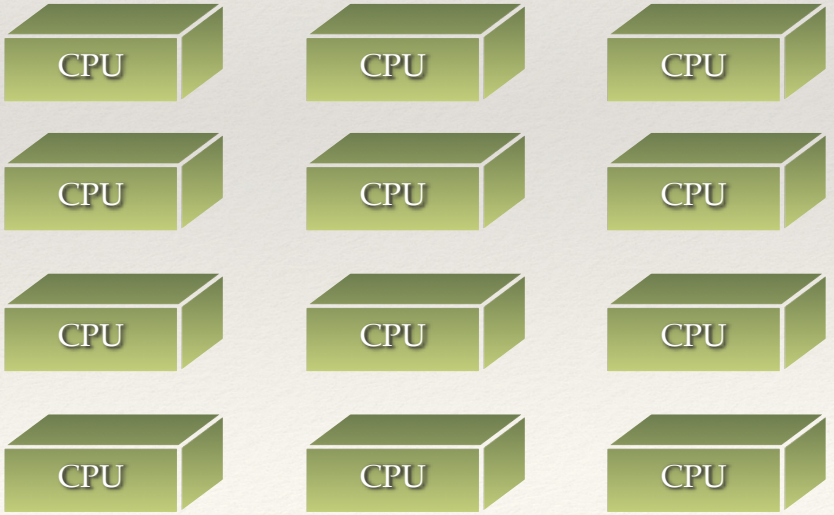


KLOE2 Experiment

Embarassing Parallel

Dati su SAN

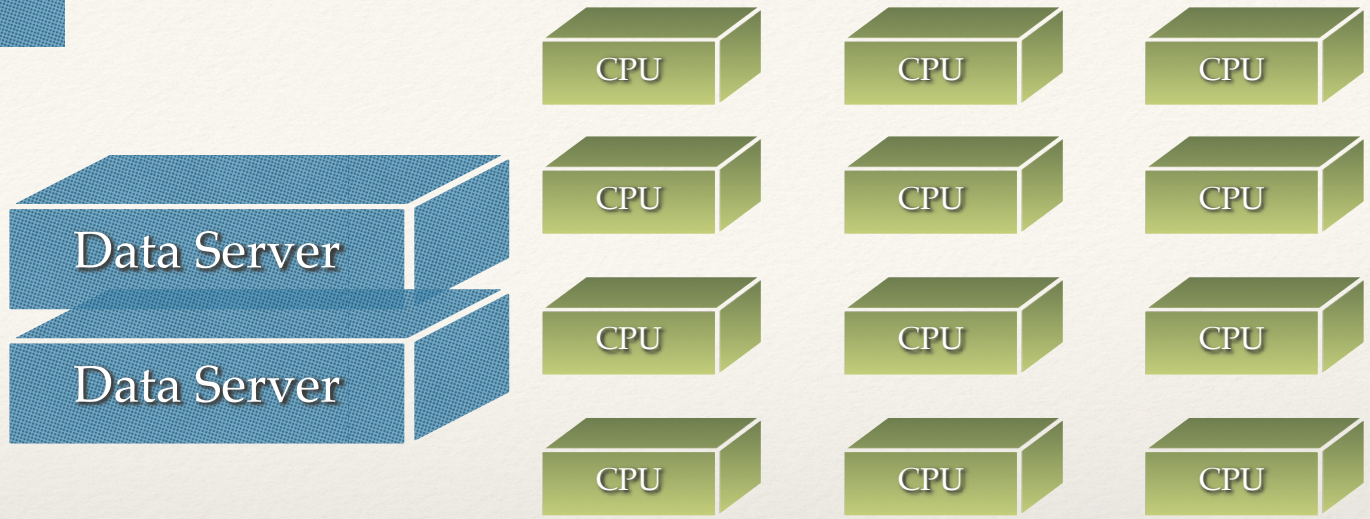
HPC Architecture



KLOE Experiment

Embarassing Parallel

Dati su NFS



Dati prelevati dal server e spediti via NFSv3 ai vari client via Ethernet

Stesso programma che gira in copia su ogni CPU e ha un suo dataset di dati in input e in output

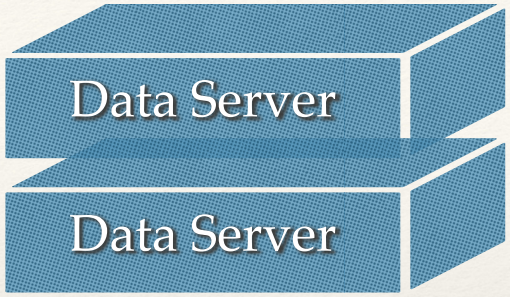
Dati elaborati vengono gestiti comunque dal server via NFSv3 ai vari client via Ethernet

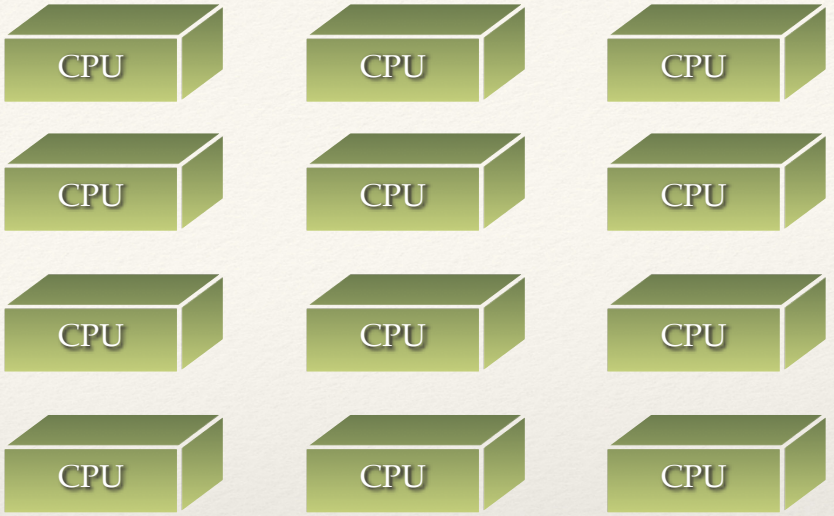
Rete sempre congestionata, tempi di latenza alti e i programmi passano molto tempo in IO-Wait State

KLOE Experiment

Embarassing Parallel

Dati su NFS



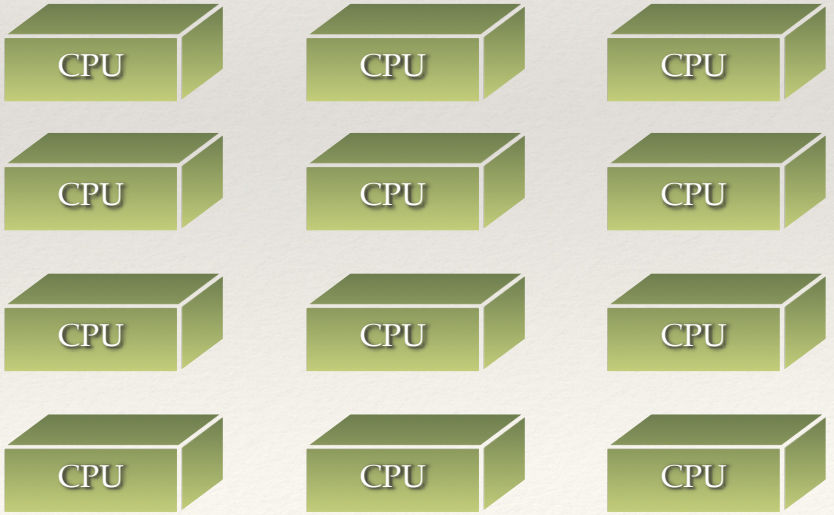


KLOE2 Experiment

Embarassing Parallel

Dati su SAN

HPC Architecture



Dati prelevati dai disk array e spediti via rete fiber channel direttamente dai vari client

Stesso programma che gira in copia su ogni CPU e ha un suo dataset di dati in input e in output

Dati elaborati vengono gestiti comunque gestiti dalla rete SAN dai vari client in modo autonomo

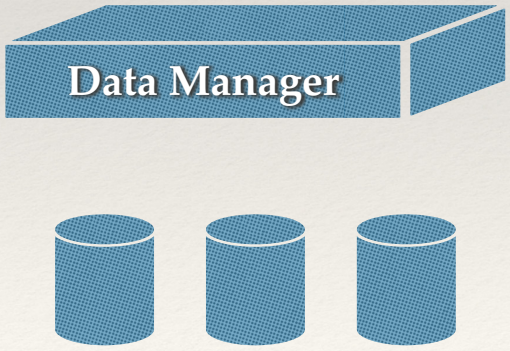
Rete mai congestionata, tempi di latenza costanti e i programmi sono sempre in zero IO-Wait state

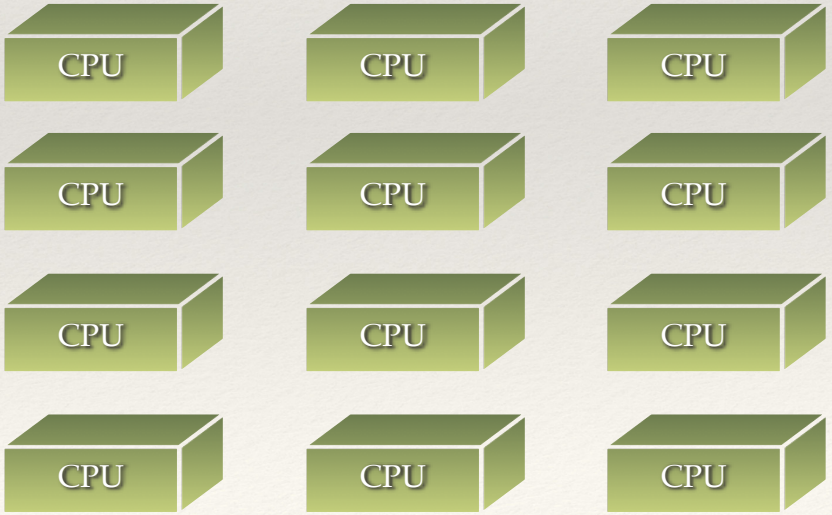
KLOE2 Experiment

Embarassing Parallel

Dati su SAN

HPC Architecture





KLOE2 impossibile senza rete fibre channel

La rete SAN ha un trhoughput altissimo rispetto a Ethernet anche a parità di velocità nominale

La rete Ethernet supporta molti protocolli

La rete fibre channel supporta un solo protocollo

La rete Ethernet NFS richiede un paio di server

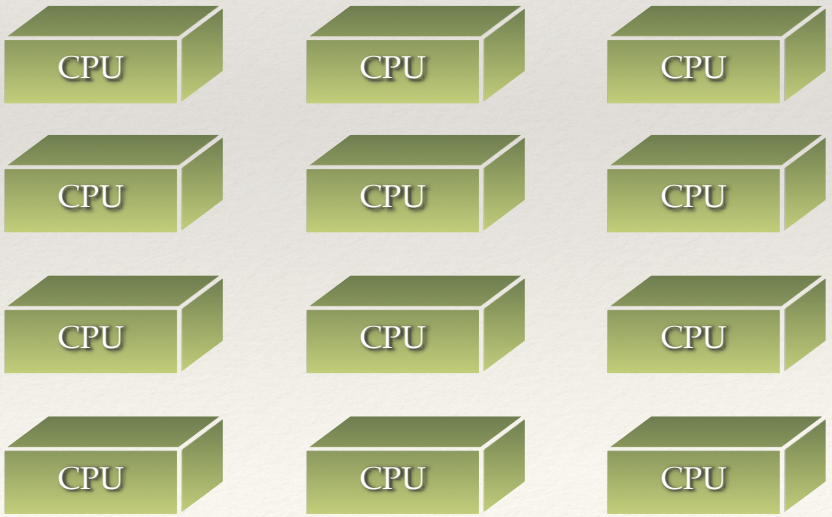
La rete fibre channel GPFS non richiede data server

KLOE2 Experiment

Embarassing Parallel

Dati su SAN

HPC Architecture



KLOE2 è stato il primo esempio di HPC nel laboratorio

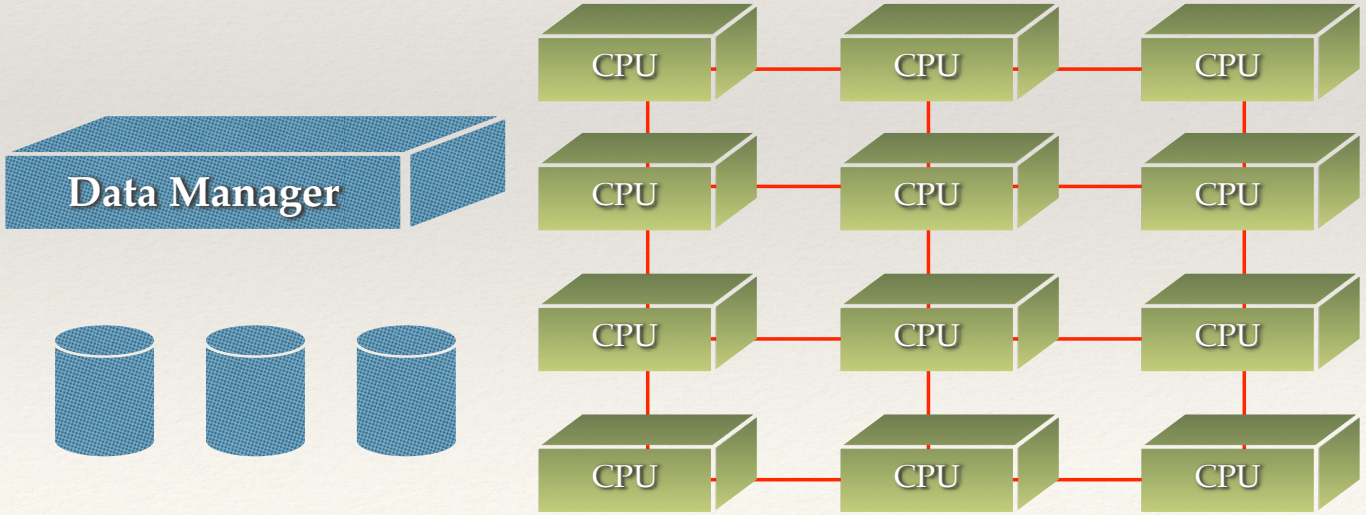
EuPRAXIA Plasma Simulation Center il secondo

La rete infiniband interconnette gli elaboratori a livello della memoria dei processori premettendo l'uso di un solo programma che gira su tutti i processori interconnessi

Terza rete per realizzare un HPC parallelo

Infiniband 100 o 200 Gbit/s

Distanza e latenza costanti tra gli elaboratori del cluster



Tre reti con protocolli differenti per l'integrazione della veicolazione dei dati sul cluster di elaboratori

Infiniband 100 o 200 Gbit/s Mellanox o NVIDIA

Connessione memory channel to memory channel dual link

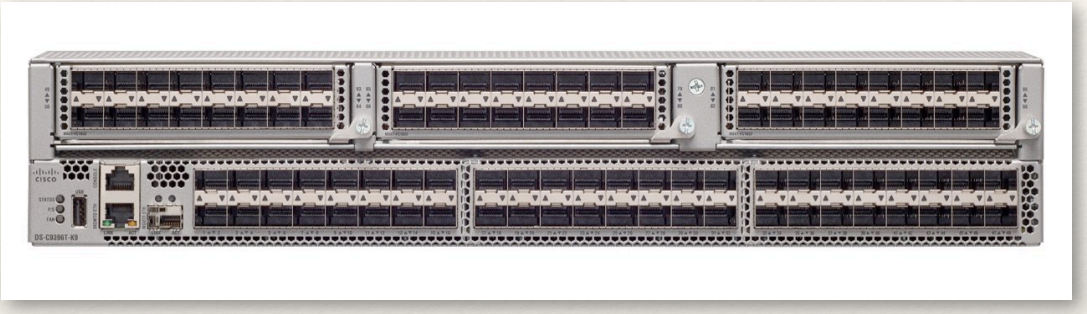
Latenza e Distanza verso ogni elemento del cluster costante

Elemento fondamentale per l'uso di programmi paralleli a memoria condivisa

Interconnette tutti i nodi del cluster

Tre reti con protocolli differenti per l'integrazione della veicolazione dei dati sul cluster di elaboratori

Fibre Channel 16/32/64 Gbit/s



Connessione per trasporto dei dati dual link multipath

Interconnette tutti i nodi del cluster

Throughput costante per i nodi grazie alla possibilità di creare trunked link multi port channel dove serveno per eliminare i colli di bottiglia

Usato con il protocollo GPFS che fornisce indipendenza di scrittura e lettura ai nodi

Tre reti con protocolli differenti per l'integrazione della veicolazione dei dati sul cluster di elaboratori

Ethernet a 10 Gbit/s

Usato solo per sottomissione job e controllo dei programmi in corso

Non c'è trasporto di dati sulla rete

Usato da gpfs solo per le richieste di metatag e per i semafori

Interconnette tutti i nodi del cluster

Infiniband 100Gbit/s

Memory to memory connection
MPI and OpenMPI protocol

Fibre Channel 32/64 Gbit/s

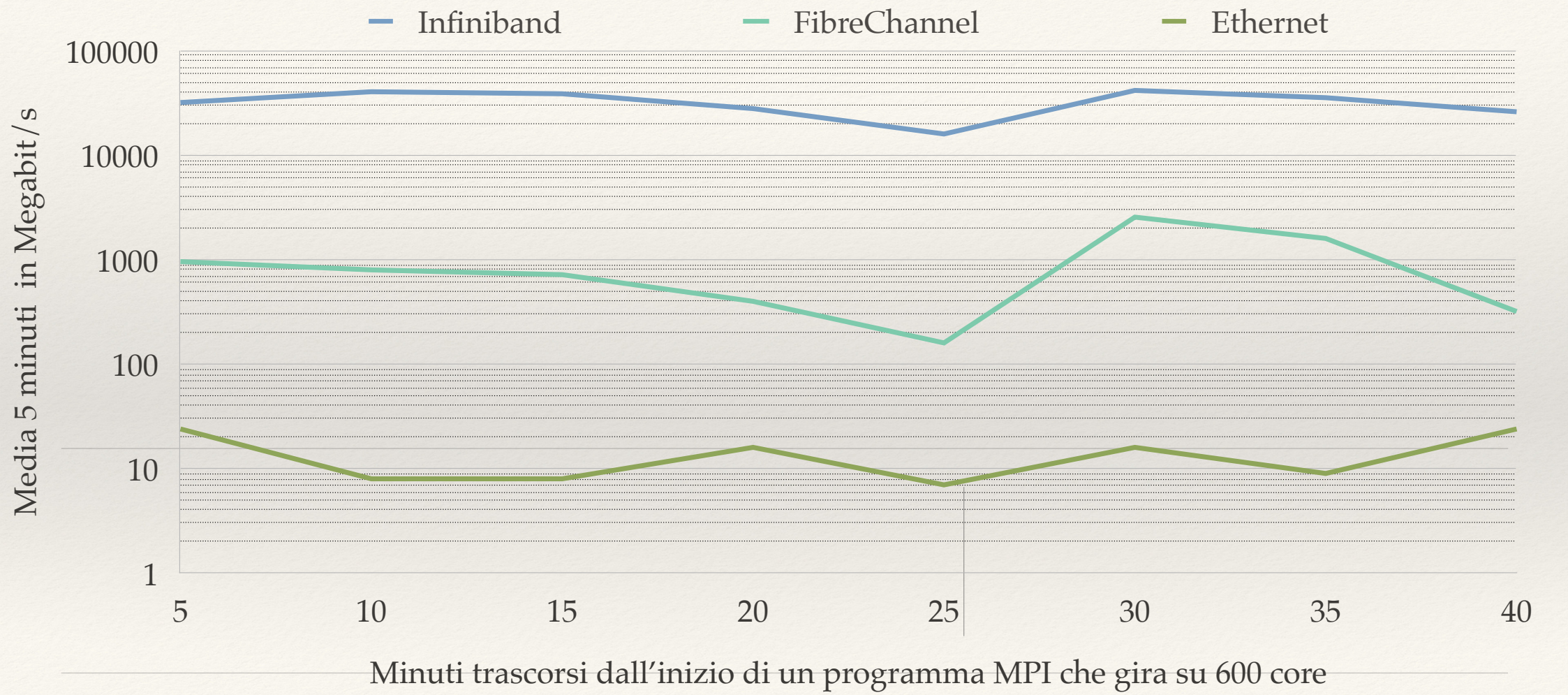
Disk Array Connection - Ruled by GPFS

Ethernet 1/10 Gbit/s full duplex

Lan connection - TCP/IP protocol

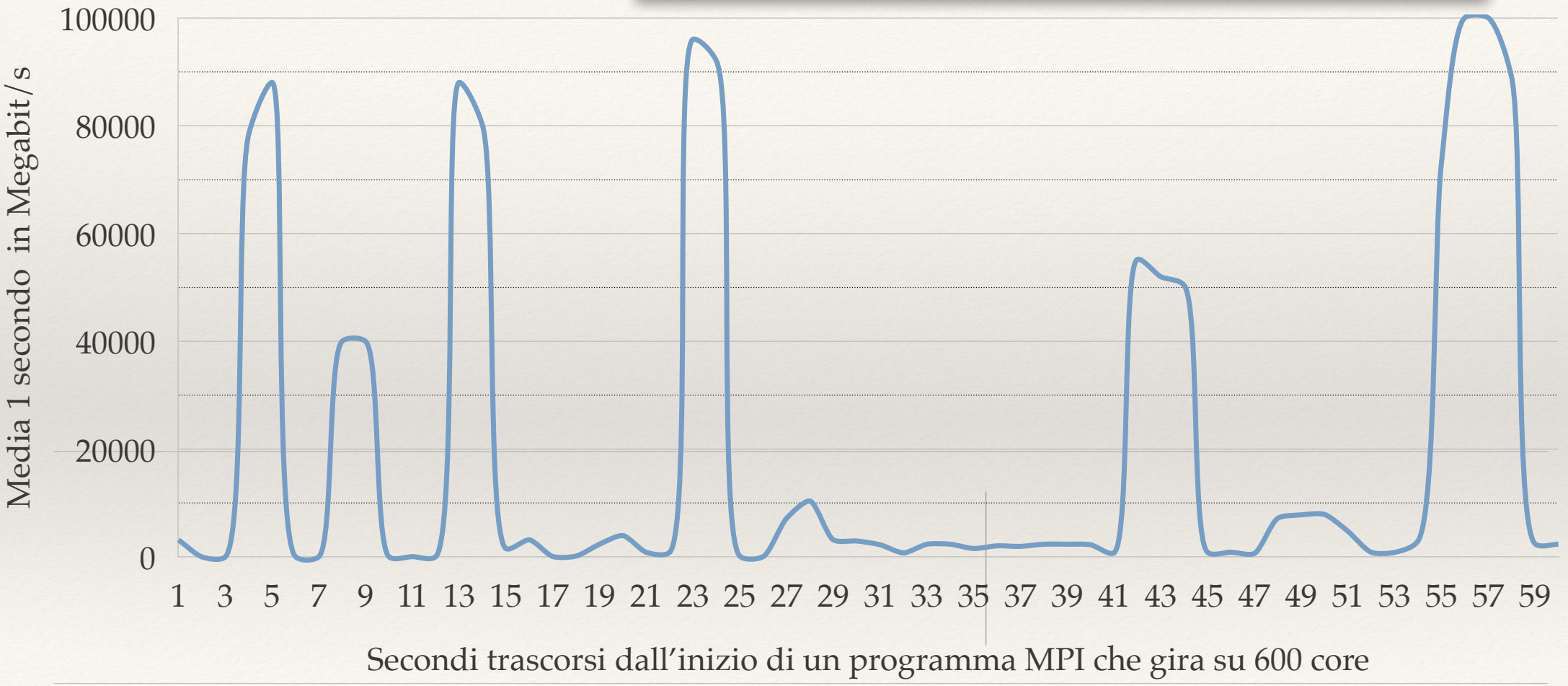
Ogni nodo del cluster dispone di tre reti di TD

Dati in transito sulle reti FC, IB e Ethernet



Rete infiniband

Sempre attiva quando un job è in corso
I valori non scendono mai sotto i 50mbit/s
Ogni tanto invece saturano la banda di 100Gbit/s



Rete Fibre Channel

Il valore di IO wait state non è mai diverso da zero su ogni nodo del cluster

Rete Ethernet

Il trasferimento di pacchetti sulla rete ethernet è ridotto a un brusio di fondo

Grazie