# DAQ innovativo per l'upgrade di LHCb

Antonio Falabella[a], Matteo Manzali[a,b]

[a] INFN - CNAF (Bologna), [b] Universià di Ferrara

IFAE 2016 - XV *Incontri di Fisica delle Alte Energie*

30 Marzo-1 Aprile - Università di Genova

# Overview



1. L'esperimento LHCb
2. Evoluzione del trigger
3. Prototipo dell'Event Building
4. Test del prototipo software
5. Scalabilità dell'EB
6. Conclusioni

# L'esperimento LHCb

DAQ innovativo per l'upgrade di LHCb

Antonio Falabella[a], Matteo Manzali[a,b]

L'esperimento LHCb

Evoluzione del trigger

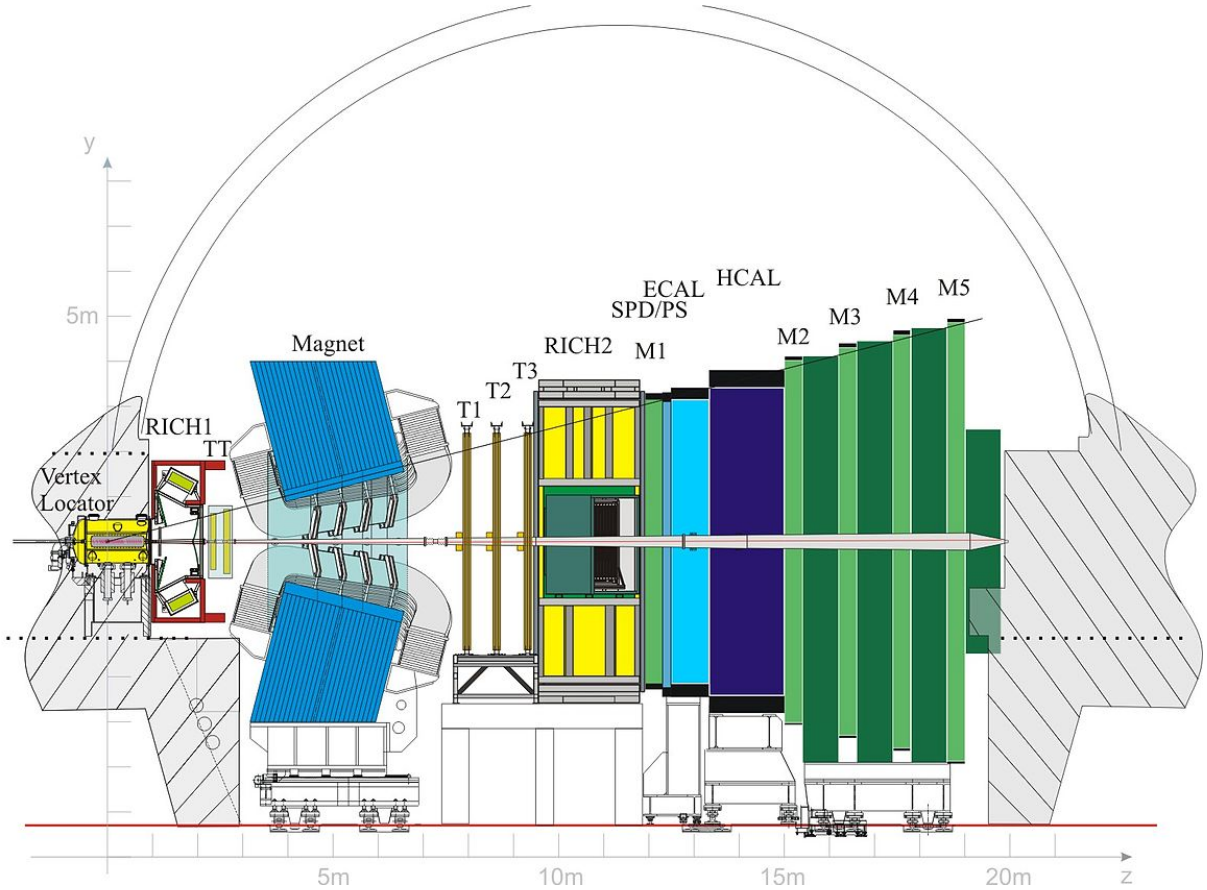
Prototipo dell'Event Building

Test del prototipo software

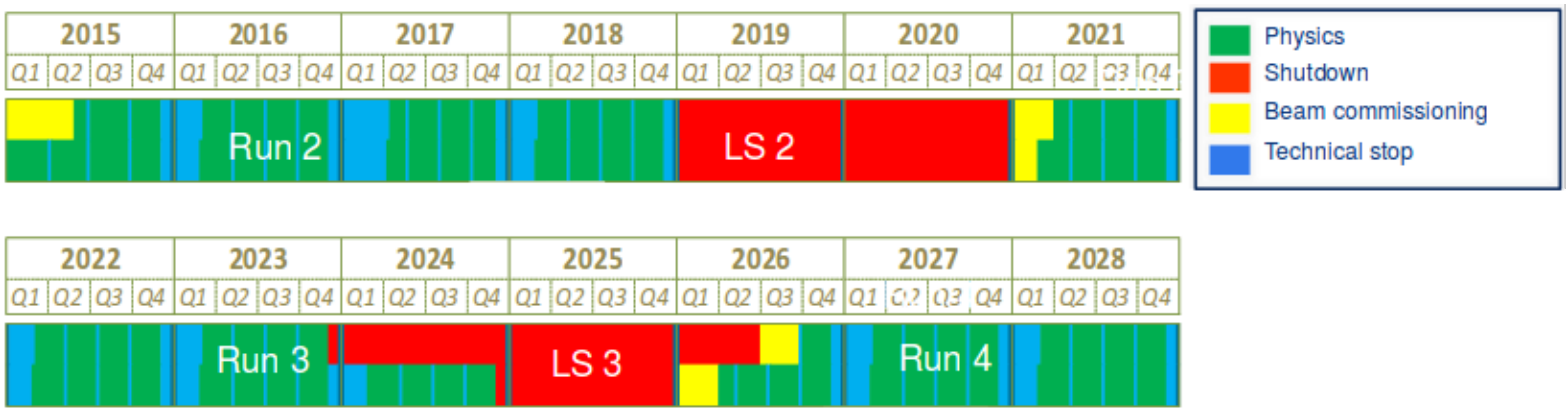
Scalabilità dell'EB

Conclusioni

- LHCb è un esperimento di fisica del sapore con lo scopo di studiare con alta precisione la violazione di CP ed i decadimenti rari degli adroni contenenti quark $b$ e $c$
- Il detector fornisce un eccellente risoluzione dei vertici e indentificazione di particelle



# Programma di presa dati



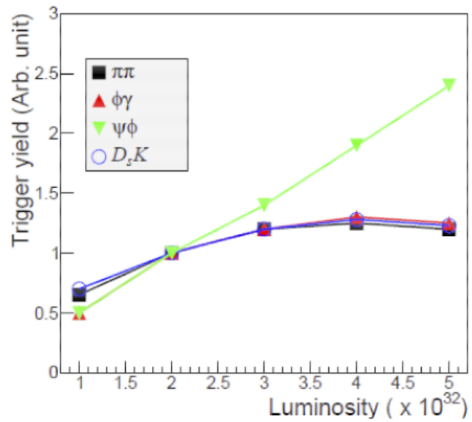

## Run II

- $1.5 \cdot 10^{34} cm^{-2} s^{-1}$
- Separazione bunch $25ns$
- Pileup $\sim 40$

## Run III

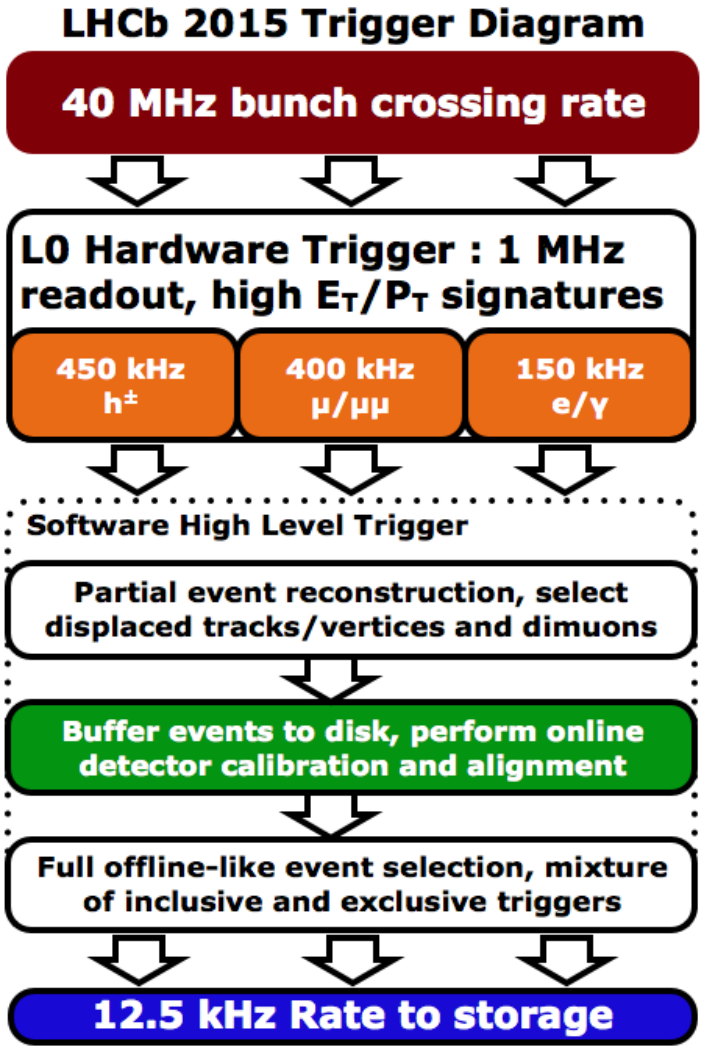
- $2.2 \cdot 10^{34} cm^{-2} s^{-1}$
- Separazione bunch $25ns$
- Pileup $\sim 60$

- La precisione di molte misure sarà limitata dalla statistica
- $\rightarrow$ aumento luminosità
- Molti canali adronici saturano l'output dell'attuale trigger



# Attuale configurazione per il trigger di LHCb

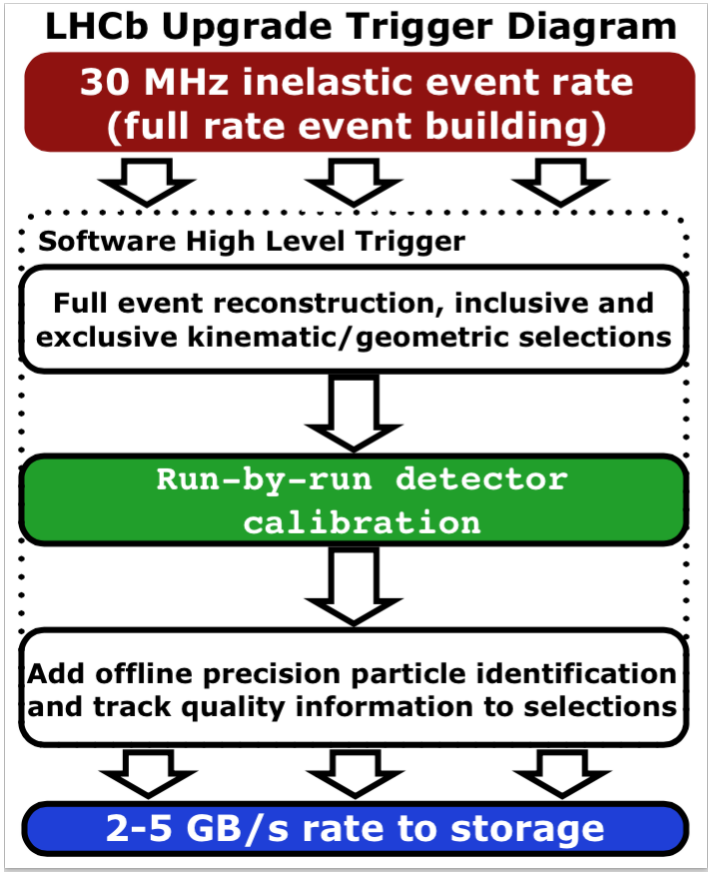
DAQ innovativo per l'upgrade di LHCb
Antonio Falabella[a], Matteo Manzali[a,b]
L'esperimento LHCb
Evoluzione del trigger
Prototipo dell'Event Building
Test del prototipo software
Scalabilità dell'EB
Conclusioni




- Sistema di trigger Run II:
- Trigger hardware L0 : riduce la frequenza da $40$ MHz a $1.1$ MHz
- Software trigger riduce ulteriormente la frequenza a $12.5$ KHz

# Configurazione del trigger di LHCb per il Run III

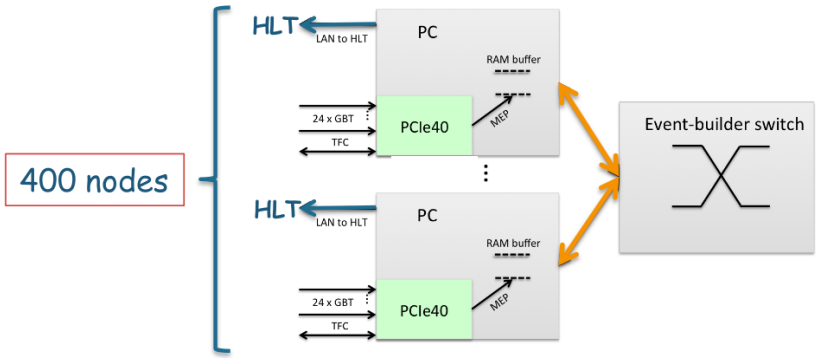
DAQ innovativo per l'upgrade di LHCb
Antonio Falabella[a], Matteo Manzali[a,b]
L'esperimento LHCb
Evoluzione del trigger
Prototipo dell'Event Building
Test del prototipo software
Scalabilità dell'EB
Conclusioni




- Sistema di trigger Run III:
- Nessun trigger hardware ($30$ MHz è la frequenza di crossing con eventi non vuoti)
- Trigger completamente software
- L'elettronica di lettura deve essere sostituita completamente

# Acquisizione dati senza trigger

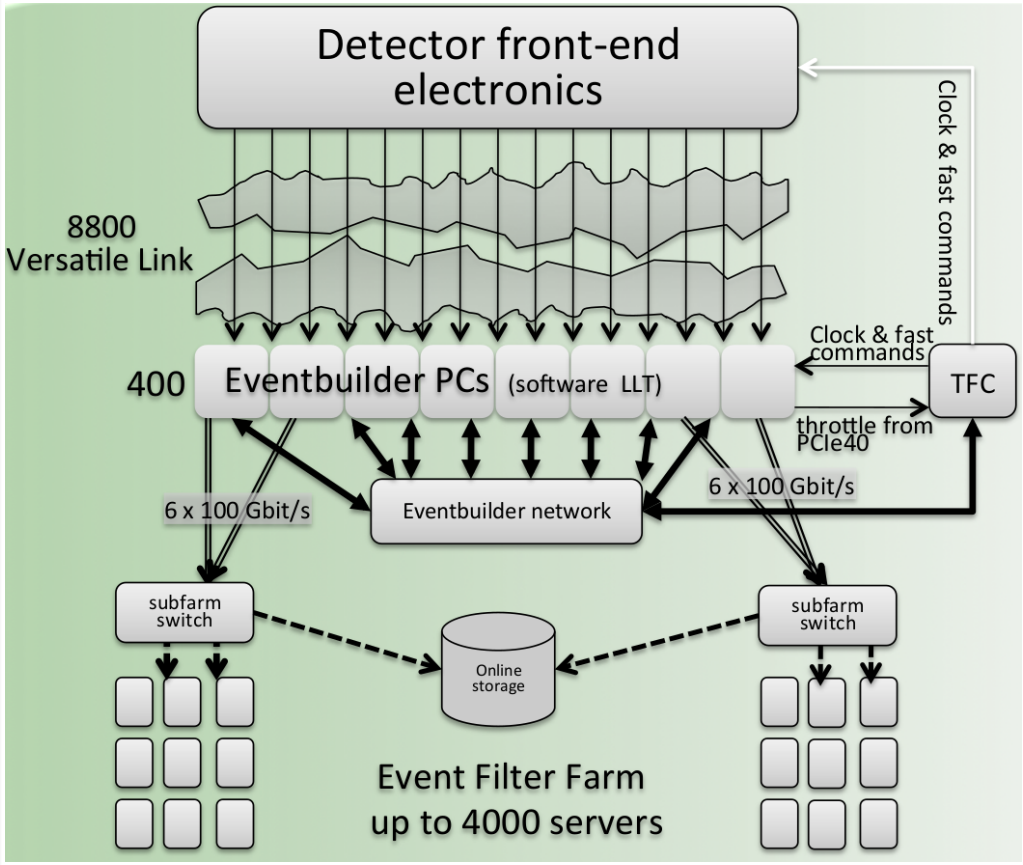
- Lettura dal rivelatore a $40\ \mathrm{MHz} \rightarrow$ composizione dell'evento a partire dalle informazioni dei sottorivelatori (*Event Building* - *EB*)
- L'event builder può essere fatto con una LAN ad alte prestazioni




- Utilizzo del protocollo PCIe versione 3 per scrivere i frammenti letti dei sottorivelatori direttamente sulla RAM dei PC dell'event builder (Multi Event Packet - MEP)

# Acquisizione dati senza trigger

DAQ innovativo per l'upgrade di LHCb
Antonio Falabella[a], Matteo Manzali[a,b]
L'esperimento LHCb
Evoluzione del trigger
Prototipo dell'Event Building
Test del prototipo software
Scalabilità dell'EB
Conclusioni




- L'EB è costituito da $\sim 400$ nodi connessi attraverso una rete ad alte prestazioni
- Ciascun nodo deve essere in grado di comunicare a $\sim 100$ Gbit/s full-duplex
- L'evento ottenuto può essere ulteriormente processato prima di essere salvato su disco

# Implementazione del DAQ

DAQ innovativo per l'upgrade di LHCb
Antonio Falabella[a], Matteo Manzali[a,b]
L'esperimento LHCb
Evoluzione del trigger
Prototipo dell'Event Building
Test del prototipo software
Scalabilità dell'EB
Conclusioni

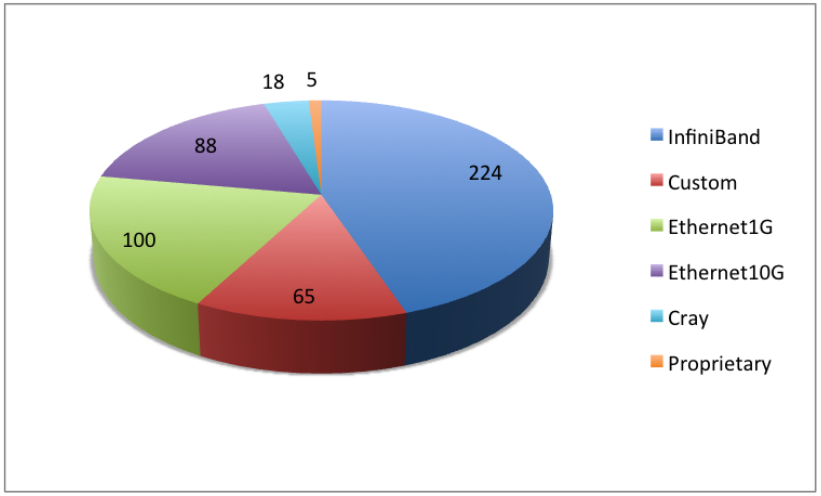
- La sfida tecnologica è quella di gestire un quantità di dati aggregata di $\sim 30$ Tbit/s (Frequenza degli eventi $\times$ Dimensione dell'evento: $30\ \text{MHz} \cdot 100\ \text{KByte}$)

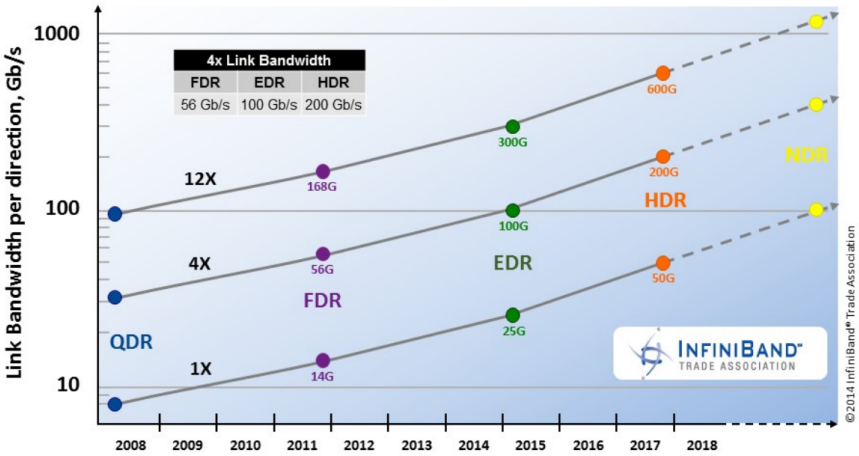
| | |
|---|---|
| Frequenza degli eventi | 30 MHz |
| Dimensione media dell'evento | 100 KBytes |
| Capacità di trasferimento dei nodi | 100 Gbit/s (PCIe 3 a 16 linee) |

- La LAN dell'EB può essere realizzata con tecnologie attualmente presenti sul mercato
- InfiniBand, 100 Gigabit Ethernet
- Questa presentazione riporta il lavoro fatto per il caso InfiniBand

# InfiniBand standard

DAQ innovativo per l'upgrade di LHCb
Antonio Falabella[a], Matteo Manzali[a,b]
L'esperimento LHCb
Evoluzione del trigger
Prototipo dell'Event Building
Test del prototipo software
Scalabilità dell'EB
Conclusioni

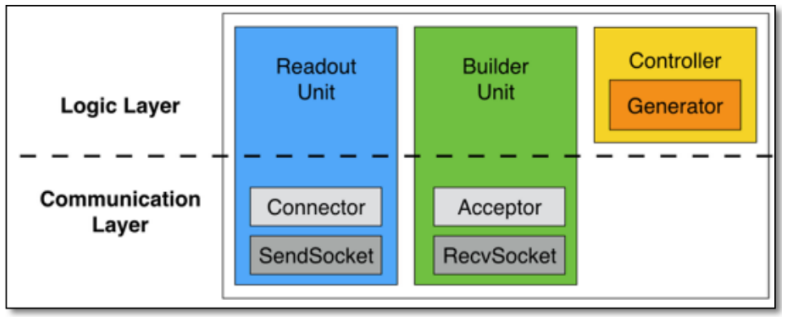



- Lo standard InfiniBand è la tecnologica più utilizzata nell'HPC
- InfiniBand garantisce alte prestazioni e basse latenze
- ... e un constante miglioramento tecnologico

# Architettura del prototipo dell'EB

- Abbiamo sviluppato un software prototipale per misurare le performance della rete basata su InfiniBand in collaborazione tra INFN e Cern (LSEB - *Large Scale Event Builder*)
- Blocchi costituenti l'EB: Readout Unit (RU), Builder Unit (BU), Generator

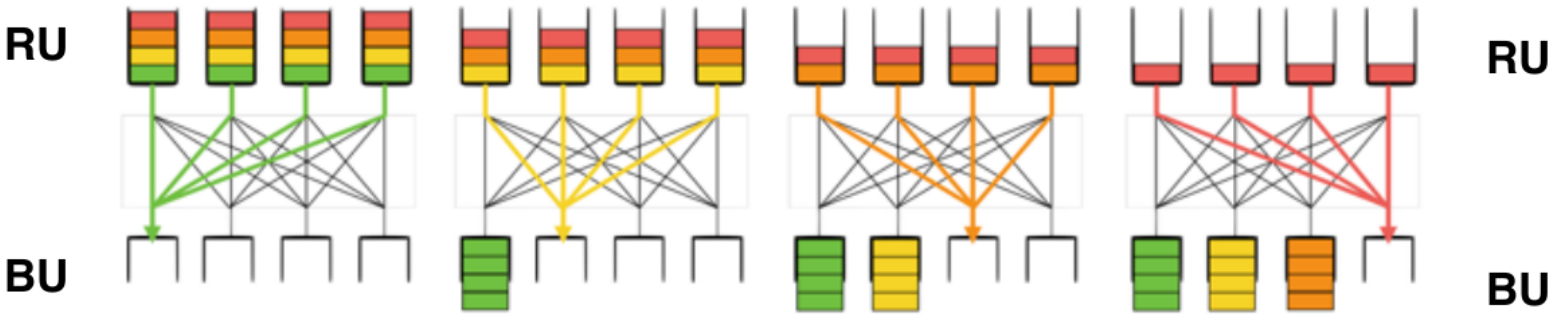



- Il generatore emula l'output delle PCIe40: genera dei frammenti di evento di dimensione opportuna
- Il generatore scrive il dato ed il corrispondente metadato direttamente nella memoria del PC che effettua la lettura (Readout Unit - RU)

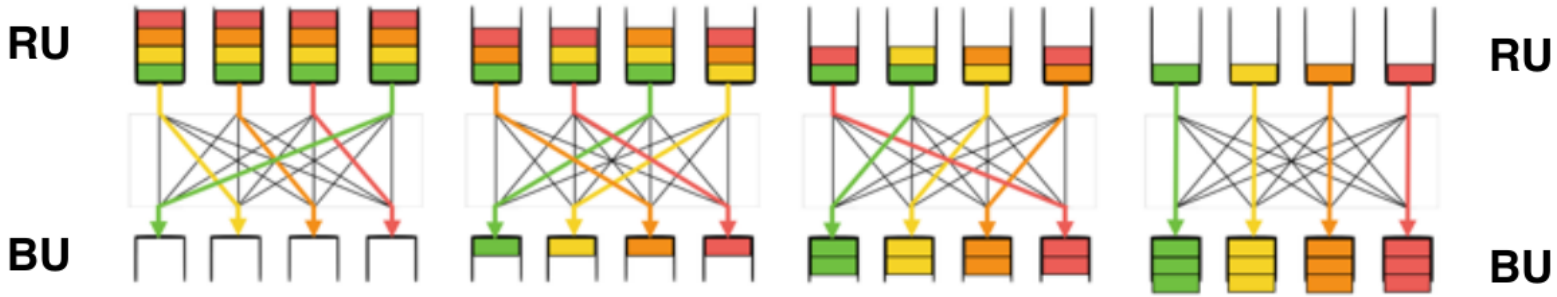
- Ciascun nodo dell'EB esegue queste tre funzioni
- A turno uno dei nodi dell'EB viene selezionato per comporre i frammenti di evento
- Ho misurato le performance su sistemi di diversa complessità e con modelli di schede di rete InfiniBand di tipo diverso

# Architettura del prototipo dell'EB

- Progettato principalmente per valutare IB, ma separando in maniera logica la parte di comunicazione dalla logica di composizione dei frammenti
- Questo garantisce la possibilità di utilizzare anche ethernet (TCP, UDP etc.)
- L'algoritmo di scelta del nodo che deve effettuare la composizione è di tipo *round robin*



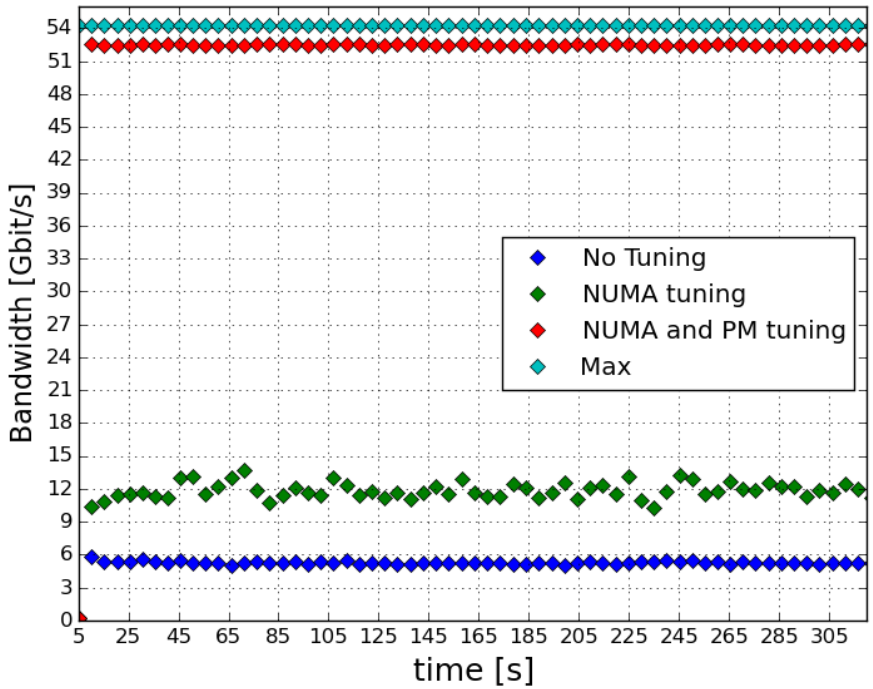

- per evitare possibili congestioni si evita la spedizione dei frammenti nello stesso istante alla stessa destinazione



# Dettagli di preparazione per i test

- Le performances di input-output possono risultare compromesse se non l'hardware non configurato correttamente. In particolare:
  - Le schede di rete devono essere collegate tramite **PCIe Gen3 (16 linee)** le precedenti versioni rappresentano un collo di bottiglia per la rete
  - Se l'architettura è tipo **NUMA** bisogna controllare lo scheduling dei processi in modo che la CPU sia sullo stesso bus della scheda di rete
  - Disabilitare il risparmio energetico e la transizione della frequenza della CPU (fonti di latenza)

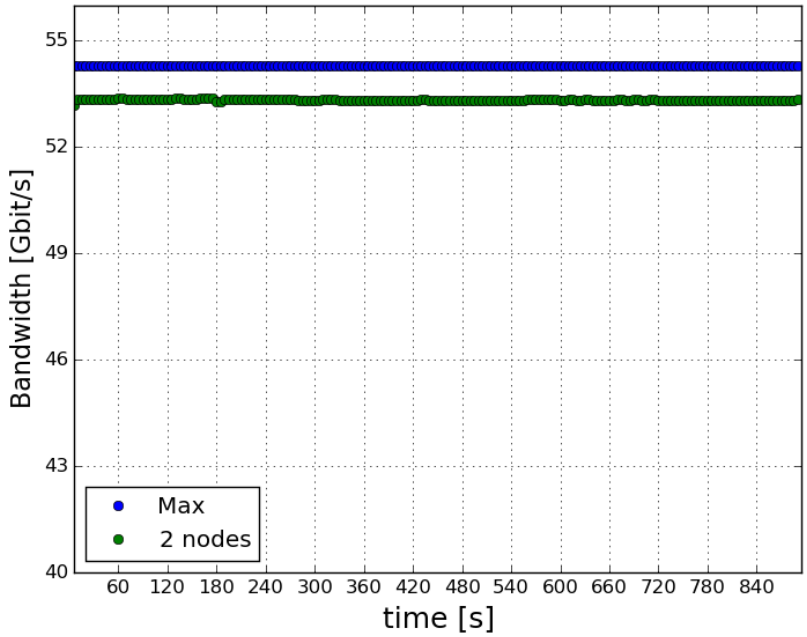
- Test effettuato al CNAF con IB FDR con due nodi collegati punto-punto



# Prestazioni dell'EB con FDR

DAQ innovativo per l'upgrade di LHCb
Antonio Falabella[a], Matteo Manzali[a,b]
L'esperimento LHCb
Evoluzione del trigger
Prototipo dell'Event Building
Test del prototipo software
Scalabilità dell'EB
Conclusioni

- Banda misurata dalle BU con un sistema costituito da due nodi con schede di rete Mellanox FDR (banda massima $54.3\ \mathrm{Gbit/s}$)

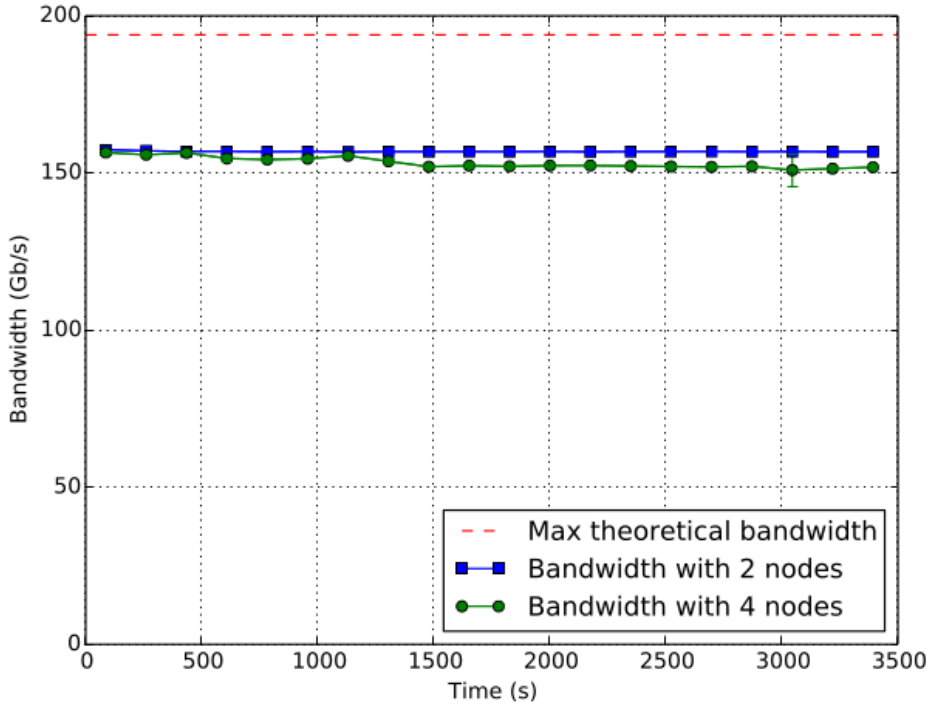



- Il valore medio misurato nell'intervallo del test (15min) è di $53.3\ \mathrm{Gbit/s}$: **98%** del massimo possibile

# Prestazioni dell'EB con EDR

DAQ innovativo per l'upgrade di LHCb
Antonio Falabella[a], Matteo Manzali[a,b]
L'esperimento LHCb
Evoluzione del trigger
Prototipo dell'Event Building
Test del prototipo software
Scalabilità dell'EB
Conclusioni

- Test effettuato con schede di rete Mellanox EDR in un cluster di 4 nodi connessi attraverso uno switch al Cern




- Banda misurata $153.9\ \mathrm{Gbit/s}$: $\sim$ **80%** della massima permessa
- Superiore alla banda richiesta per l'EB

# Scalabilità del software dell'EB

DAQ innovativo per l'upgrade di LHCb
Antonio Falabella[a], Matteo Manzali[a,b]
L'esperimento LHCb
Evoluzione del trigger
Prototipo dell'Event Building
Test del prototipo software
Scalabilità dell'EB
Conclusioni

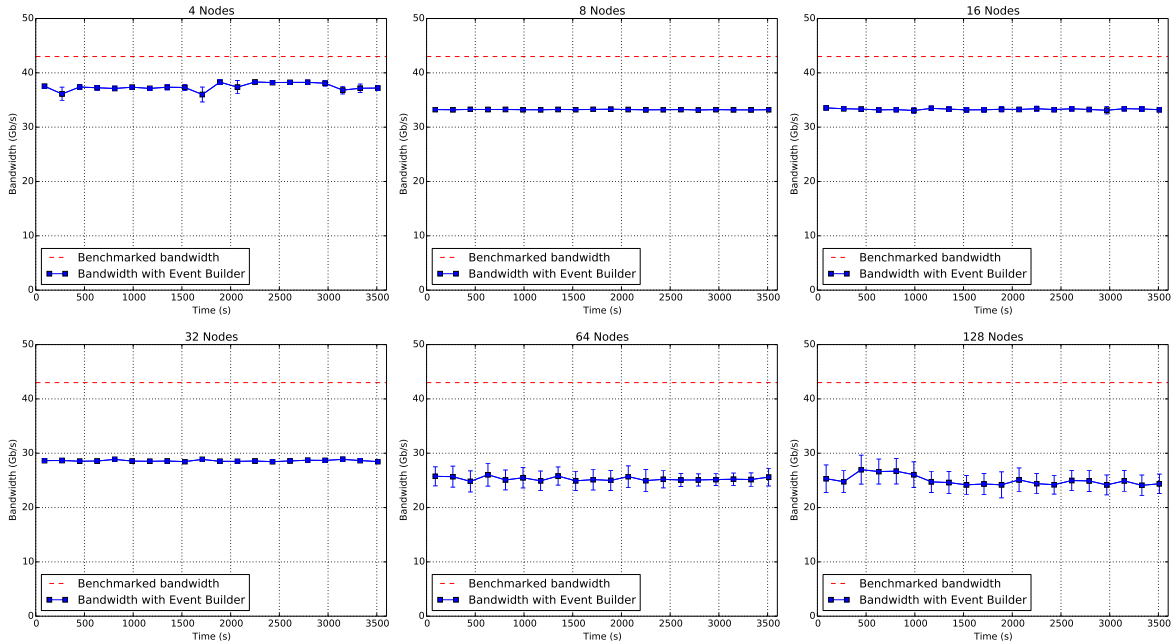
- Test di scalabilità effettuati al cluster Galileo del consorzio CINECA ▸ Link

| | |
|---|---|
| Nodi | 516 |
| Processors | 2 8-core Intel Haswell $2.40\ \mathrm{GHz}$ per node |
| RAM | 128 GB/node, 8 GB/core |
| Rete | InfiniBand con switch QDR |

- Le dimensioni del cluster sono simili alle dimensioni della LAN prevista per l'EB

# Prestazioni dell'EB

- Blue: Banda aggregata nelle due direzioni dell'EB (massimo $43$ Gbit/s)



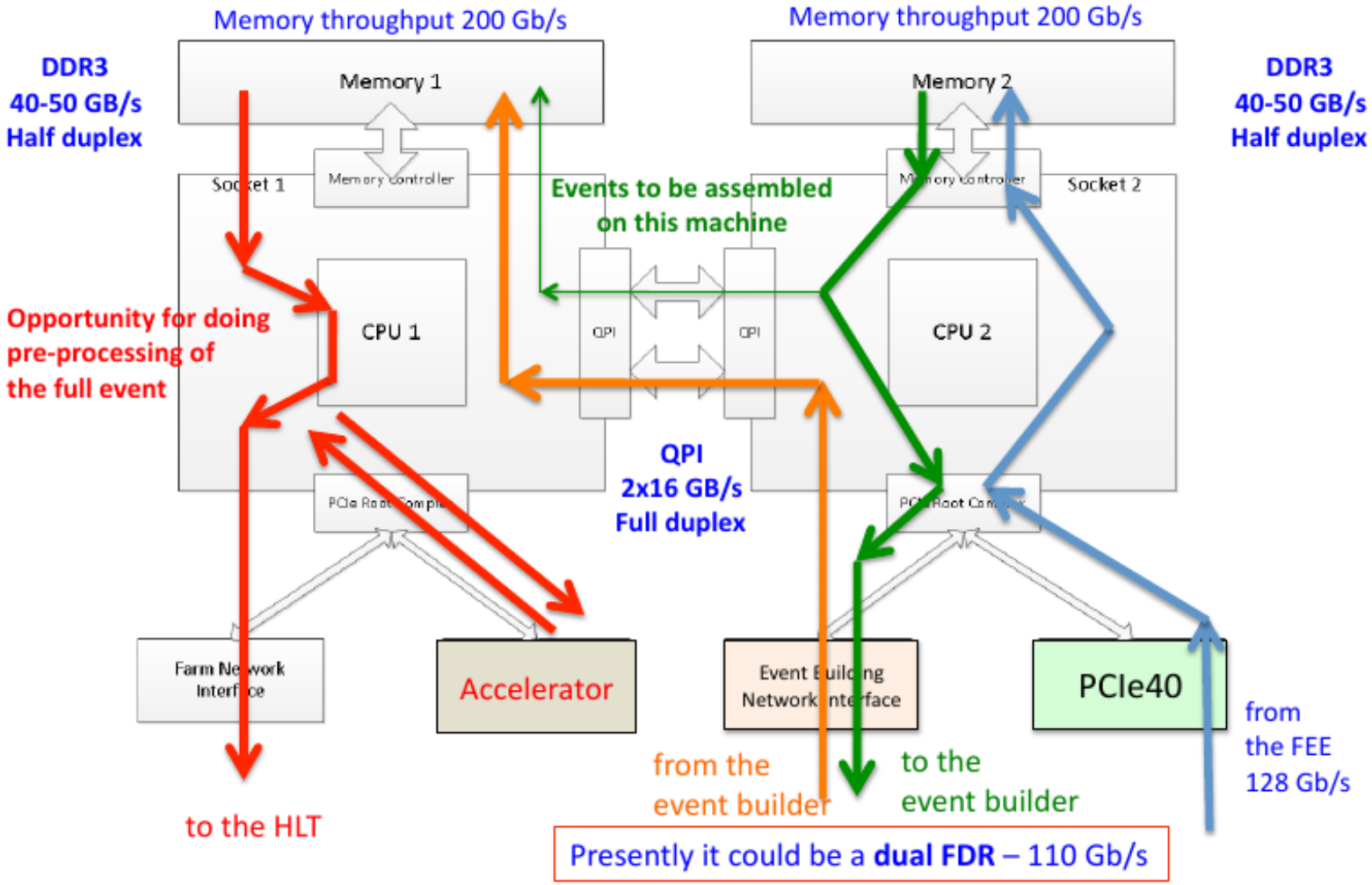

- L'EB funziona correttamente anche alla complessità di **128** nodi: $60\%$ della banda possibile
- Alcune limitazioni intrinseche :
  - Il cluster è in produzione, quindu numerosi altri processi che utilizzano la rete.
  - **nessun controllo sul risparmio energetico e la frequenza delle CPU**

# Conclusioni

- L'Upgrade dell'esperimento LHCb per il Run III è una sfida tecnologica in particolare per il DAQ a $40\mathrm{MHz}$
- Un trigger solo software può essere implementato utilizzando un EB con LAN basata su InfiniBand
- Abbiamo sviluppato un software per valutare questa tecnologia in collaborazione tra INFN e Cern
- Ho testato diversi tipi di schede di rete ed effettuato un test di scalabilità mostrando la fattibilità di questa soluzione, e la robustezza del nostro software
- Sviluppi ulteriori:
  - Implementare un meccanismo di tolleranza al fallimento di un nodo
  - effettuare test su scala più ampia

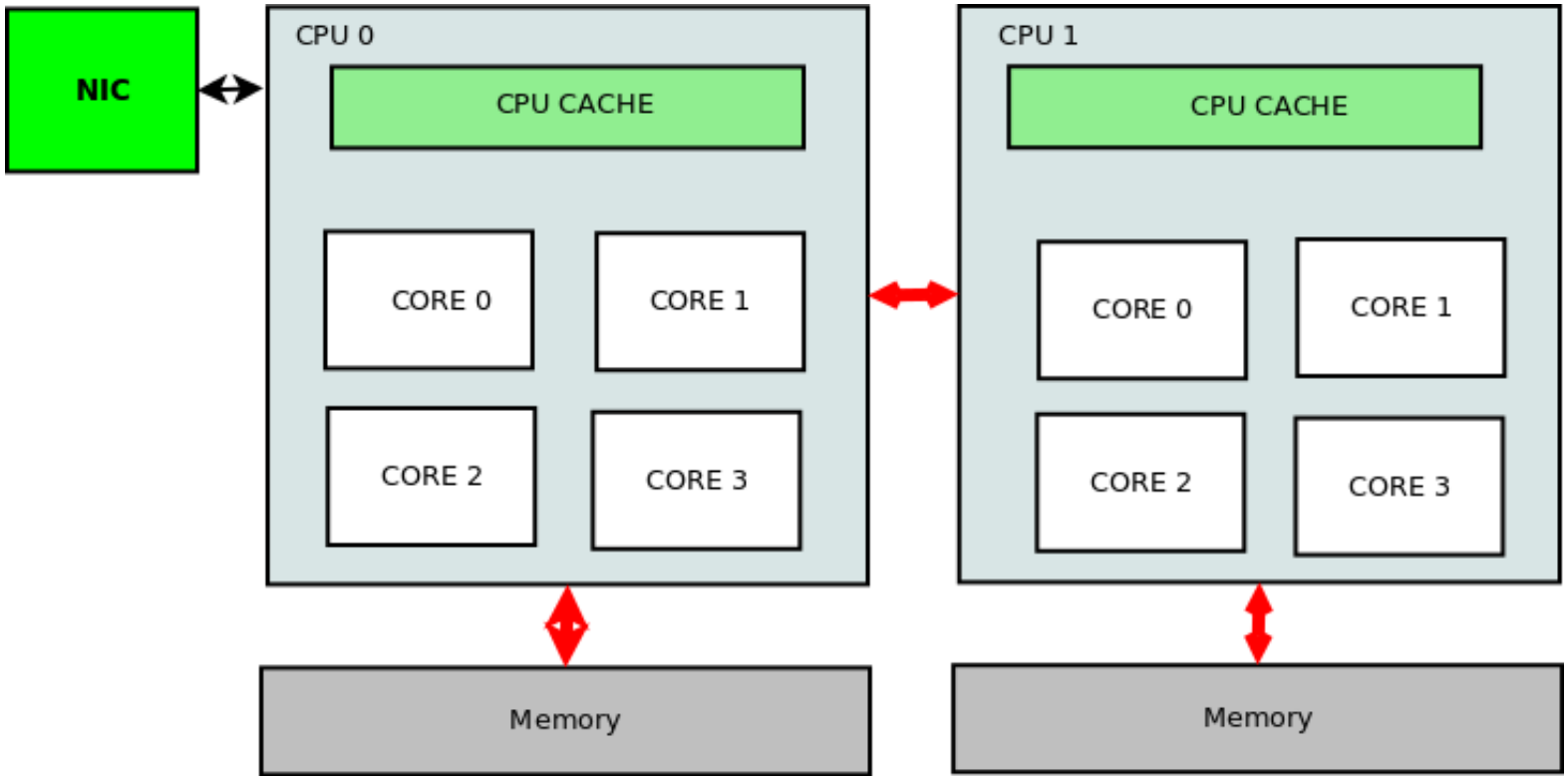
# Backup

DAQ innovativo per l'upgrade di LHCb
Antonio Falabella[a], Matteo Manzali[a,b]
L'esperimento LHCb
Evoluzione del trigger
Prototipo dell'Event Building
Test del prototipo software
Scalabilità dell'EB
Conclusioni



# Numa - Non Uniform Memory Access

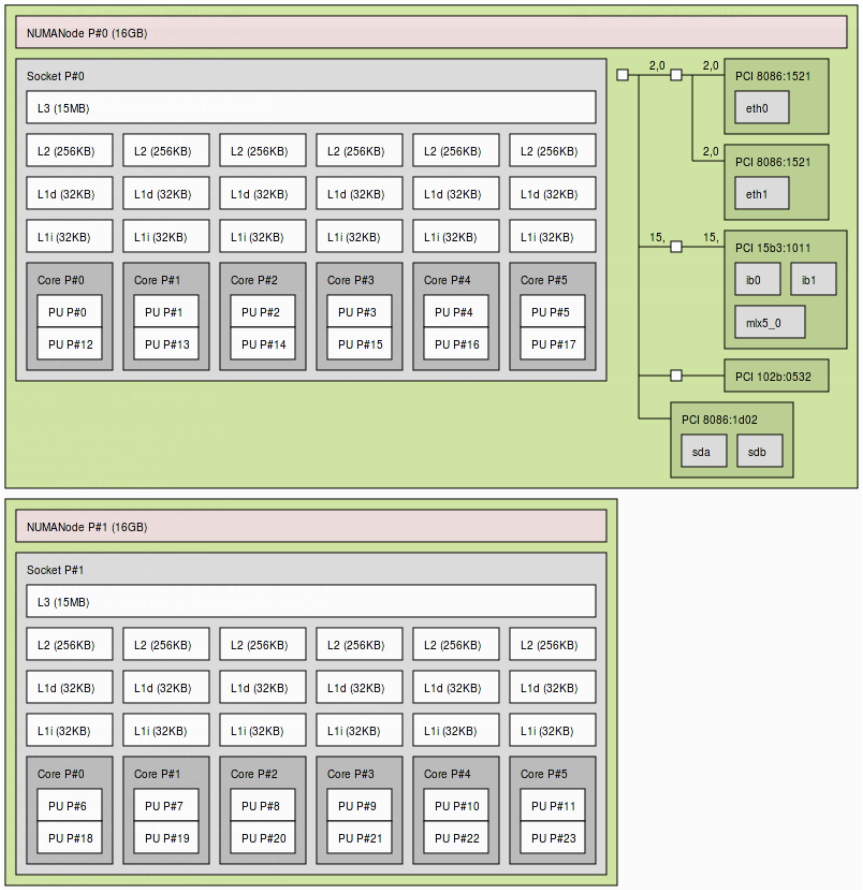
- Architettura di memoria per sistemi multiuprocessore
- Nel schemo ogni CPU (nodo NUMA)ha il suo banco di memoria locale
- Una CPU ha accesso più veloce alla sua memoria locale




- L'interfaccia di rete è collegata ad una delle CPU
- Se la CPU1 tenta di accedere all'interfaccia di rete soffrirà di una latenza maggiore rispetto alla CPU0

# Numa topology

Con *lstopo*, parte del pacchetto *hwloc* è possibile produrre schemi per la CPU/Cache/Memoria




Esempio con 2 macchine

- Le interfacce di rete sono ib0 e ib1 sono connesse ad uno delle due CPU
- Se i frammenti sono gestiti da un processo sui core a 6 a 11 vi sarà una latenza aggiuntiva

# Problematiche di gestione energetica

DAQ innovativo per l'upgrade di LHCb

Antonio Falabella[a], Matteo Manzali[a,b]

L'esperimento LHCb

Evoluzione del trigger

Prototipo dell'Event Building

Test del prototipo software

Scalabilità dell'EB

Conclusioni

- Gli stati di risparmio energetico della CPU detti C-states possono essere critici per le prestazioni di input-output
- Lo stato C-0 corrisponde a nessun risparmio energetico. C-X con X¿0 sono stati di risparmio energetico via via maggiore
- La transizione tra questi stati è una fonte di latenza per le applicazioni che fanno molto IO