New Cineca Tier-0 Marconi

Carlo Cavazzoni

Cineca

High level system Characteristics

Tender proposal

Partition	Installation	CPU	# nodes	# of Racks	Power
A1 – Broadwell (2.1PFlops)	April 2016	E5-2697 v4	1512	25	700KW
A2 - Knight Landing (11 Pflops)	September 2016	KNL	3600	50	1300KW
A3 – Skylake (4.5PFlops)	June 2017	E5-2680 v5	1512	25	700KW

Network: Intel OmniPath

Marconi - Compute

Partizione A1

1512 Lenovo NeXtScale Server -> 2PFlops
processore Intel E5-2697 v4 Broadwell
18 cores @ 2.3GHz.
dual socket node: 36 core e 128GByte / nodo

Partizione A2

3600 server Intel AdamPass -> 11PFlops
processore Intel PHI code name Knight Landing
68 cores @ 1.4GHz.
single socket node: 96GByte DDR4 + 16GByte MCDRAM

Partizione A3

1512 Lenovo Stark Server -> 4.5PFlops
processore Intel E5-2680v5 SkyLake
20 cores @ 2.??GHz
dual socket node: 40 core e 196GByte /nodo

System layout

- CINECA Floor Plan

System A1:

- Mgmt (1x)
- Storage-Nodes (1x)
- GSS (4 x)
- OPA (5 x)
- BDW (21 x)

System A2:

- KNL (51 x)

System A3:

- SKL (21 x)

Marconi - Network

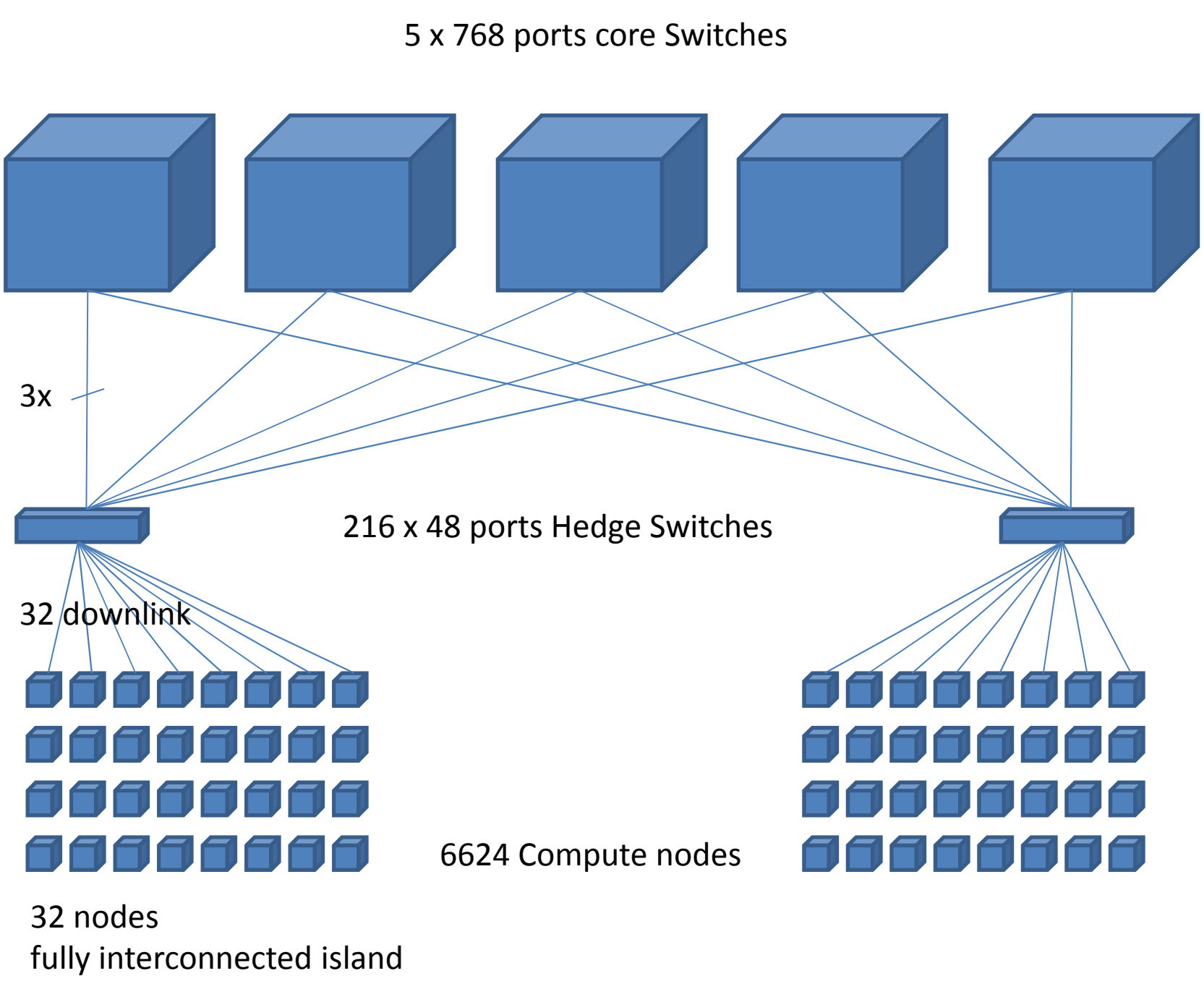
Network type: new Intel Omnipath
Largest Omnipath cluster of the world

Network topology: Fat-tree
2:1 oversubscription
tapering at the level of the core switches only

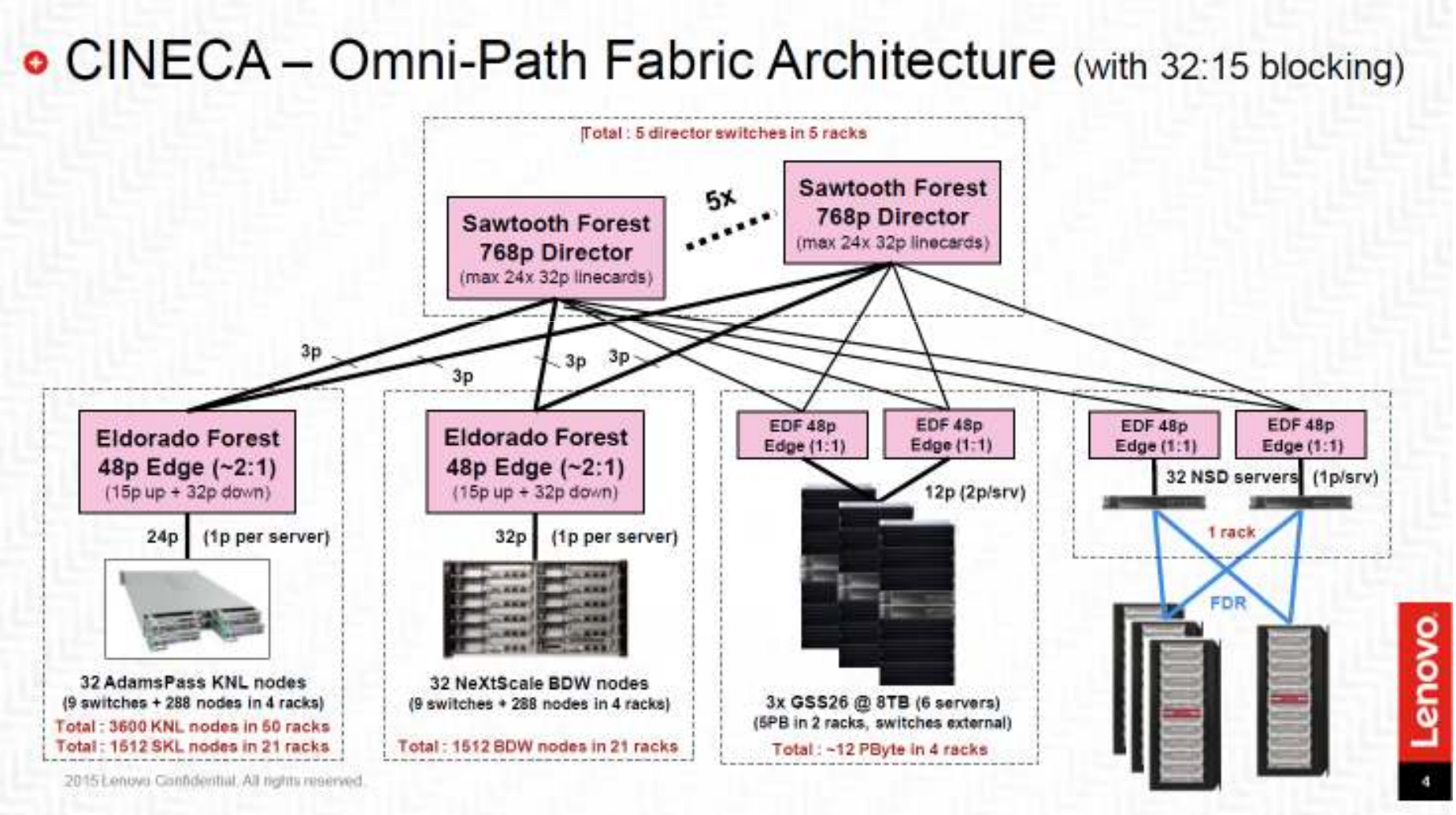
Core Switches: 5 x OPA Core Switch “Sawtooth Forest”
768 ports each

Hdge Switch: 216 OPA Edge Switch “Eldorado Forest”
48 ports each

Maximum system configuration:
5(opa) x 768(ports) x 2(tapering) -> 7680 servers


5 x 768 ports core Switches
3x
216 x 48 ports Hedge Switches
32 downlink
6624 Compute nodes
32 nodes
fully interconnected island

System Layout



Marconi - Storage

Storage system:

6 x Lenovo GSS-26 Storage

Storage capacity:

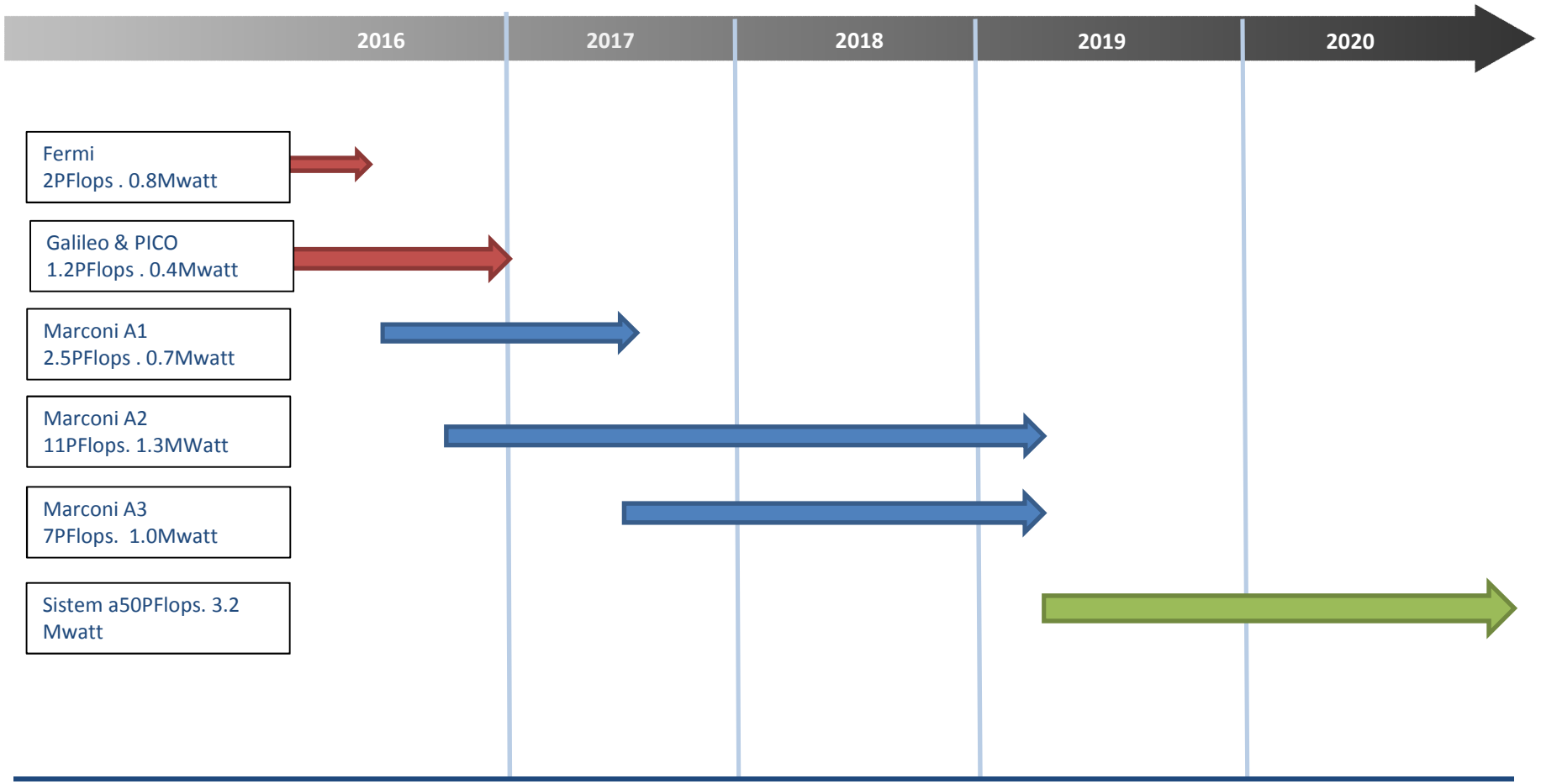
17PB(raw capacity)

Storage bandwidth:

100GByte/sec (sustained)

Storage network:

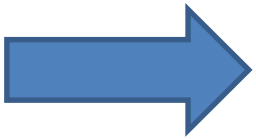
Intel Omnipath (directly attached to the OPA switchs)



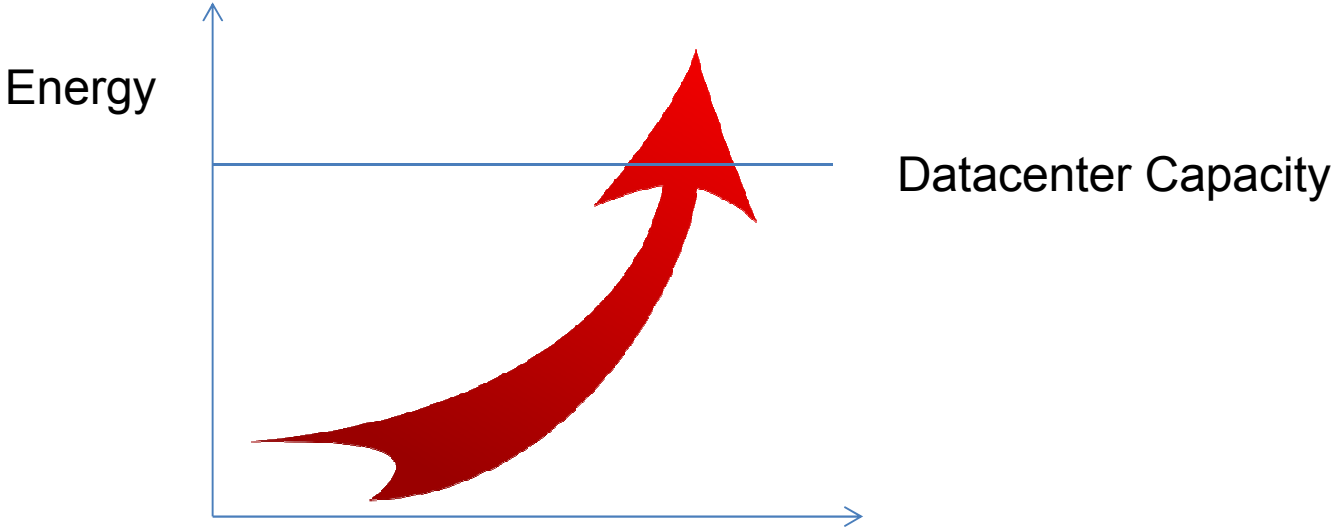
1.2Mwatt	2.4Mwatt	2.3Mwatt	2.3Mwatt	3.2Mwatt
50 rack	120 rack	120 rack	120 rack	150 rack
100mq	240mq	240mq	240mq	300mq

Energy trends

"traditional" RISK and CISC chips are designed for maximum performance for all possible workloads



A lot of silicon to maximize single thread performace



Change of paradigm

New chips designed for maximum performance in a small set of workloads

Simple functional units, poor single thread performance, but maximum throughput