TECHNISCHE UNIVERSITÄT DARMSTADT

Large-Scale Parallel Computing

Prof. Dr. Felix Wolf

PARALLEL ARCHITECTURES

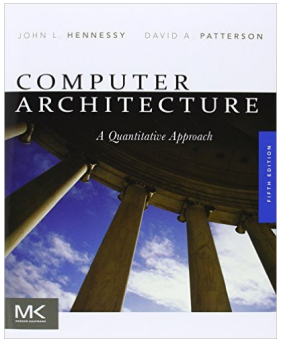
Outline

- Classification
- Memory architecture
- Interconnection networks
- Example: IBM BlueGene/Q

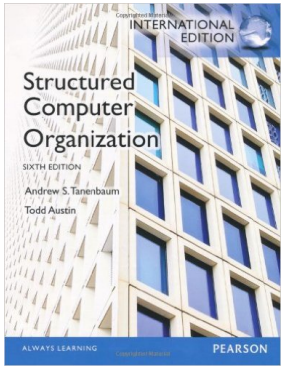
Literature

TECHNISCHE UNIVERSITÄT DARMSTADT

- John L. Hennessy, David A. Patterson: Computer Architecture: A Quantitative Approach, 5th edition, Morgan Kaufmann, 2011



- Andrew S. Tanenbaum, Todd Austin: Structured Computer Organization, 6th edition, Pearson, 2013



Taxonomies

TECHNISCHE UNIVERSITÄT DARMSTADT

- Number of instruction streams vs. number of data streams
- Memory architecture
- Network architecture
- Degree of heterogeneity
- Degree of customization
- Size of nodes in relation to number of nodes

Flynn's classification [1966]

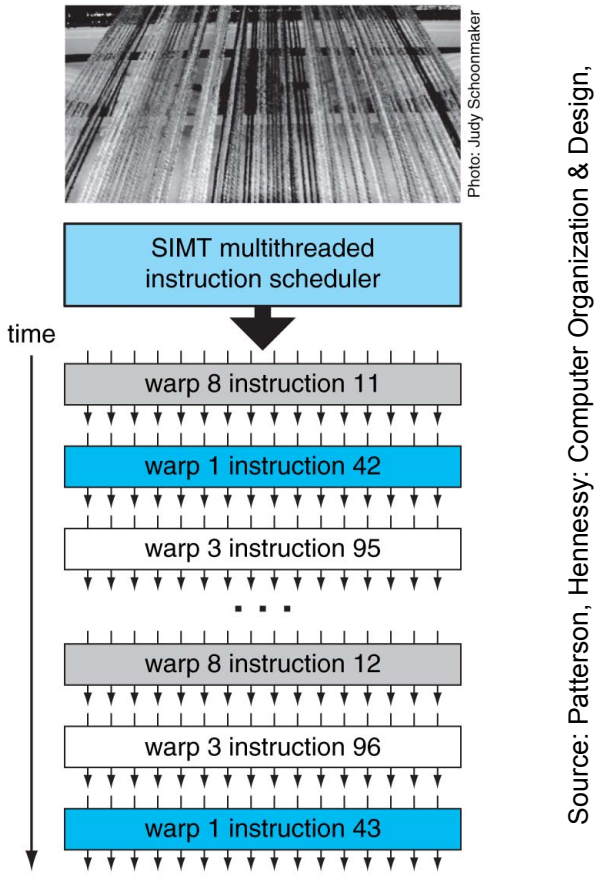
TECHNISCHE UNIVERSITÄT DARMSTADT

- Single instruction stream, single data stream
  - Classical uniprocessor
- Single instruction stream, multiple data streams
  - Same instruction is executed by multiple processors using different data streams
  - Data parallelism
  - Examples: SIMD extensions for multimedia, vector processors
- Multiple instruction streams, single data stream
  - No commercial multiprocessor of this type ever built
- Multiple instruction streams, multiple data streams
  - Each processor fetches its own instructions and operates on its own data
  - Thread-level parallelism

Single-instruction multiple threads (SIMT)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Creates, manages, schedules, and executes threads in groups of parallel threads called warps
- At each instruction issue time, SIMT instruction unit
  - Selects warp that is ready to execute its next instruction
  - Broadcasts instruction to all active threads of that warp
- Individual threads may be inactive to do independent branching



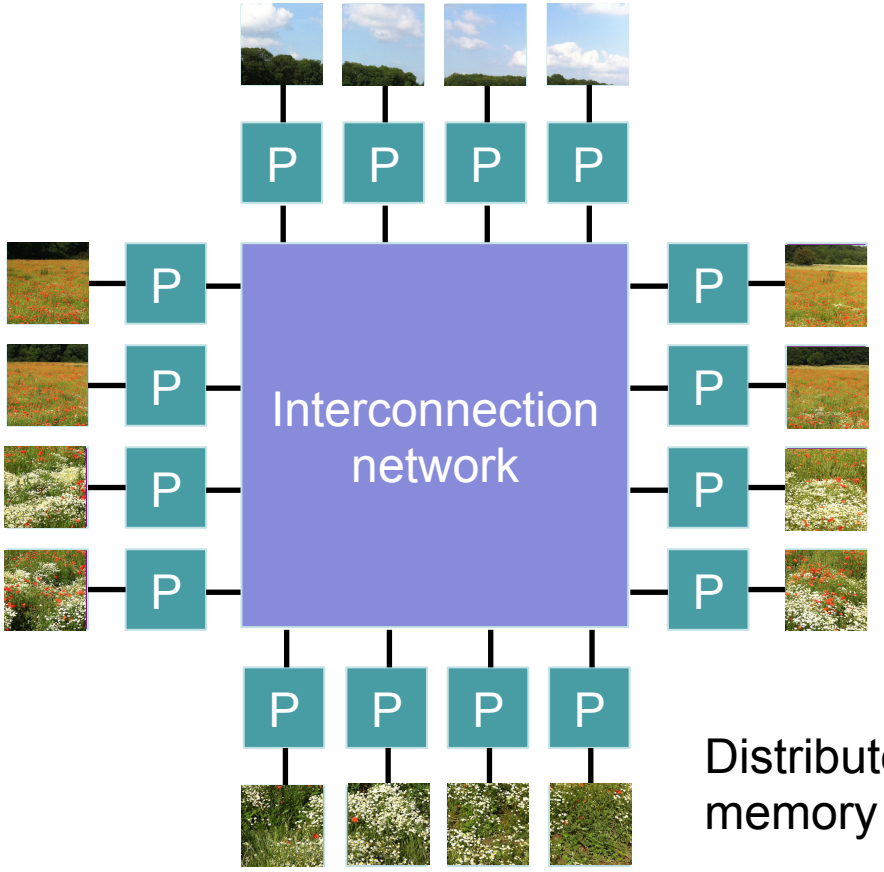
Source: Patterson, Hennessy: Computer Organization & Design, 4th edition, Morgan Kaufmann

MIMD

TECHNISCHE UNIVERSITÄT DARMSTADT

- Architecture of choice for general-purpose multiprocessors
- Offers high degree of flexibility
  - High performance for one application or multi-programmed multiprocessor
- Can take advantage of off-the-shelf processors
- Popular execution model - Single Program Multiple Data (SPMD)
  - The same program is executed in parallel with each instance having a potentially different control flow

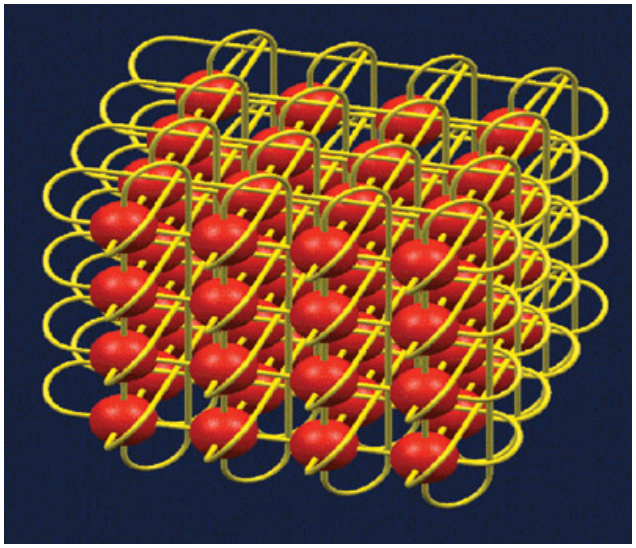
Memory architecture



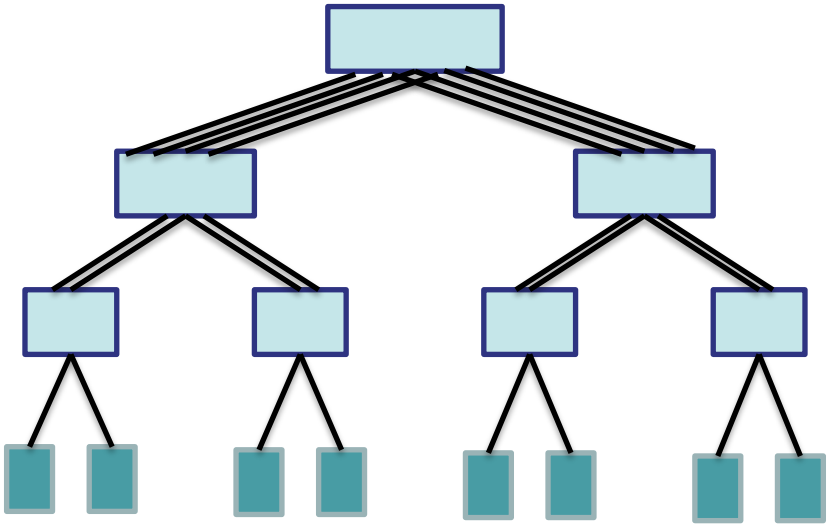
Distributed memory

Popular network architectures for distributed memory systems

TECHNISCHE UNIVERSITÄT DARMSTADT



Torus
(distributed switched network)

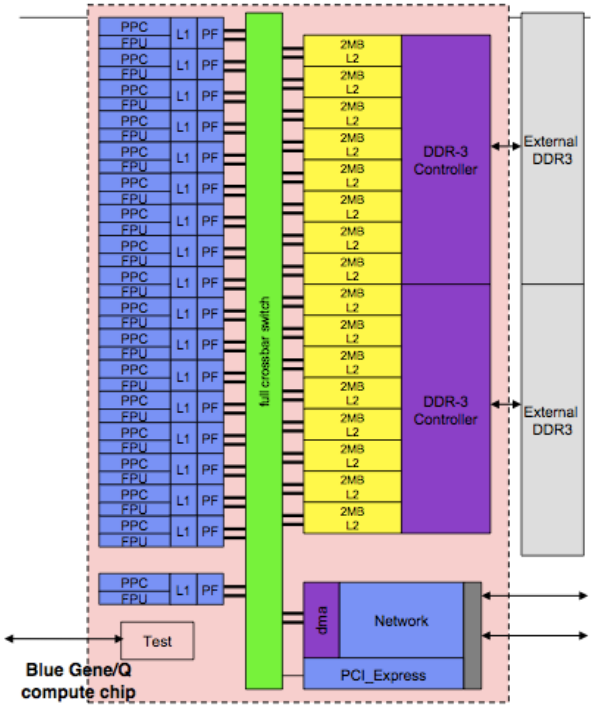


Fat tree
(centralized switched network)

Degree of heterogeneity

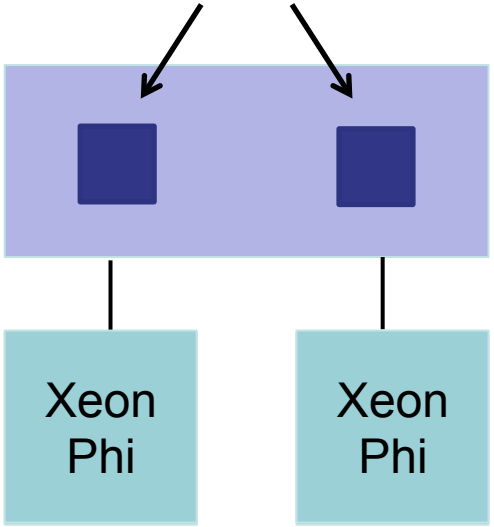
TECHNISCHE UNIVERSITÄT DARMSTADT

Homogeneous node architecture



BlueGene/Q chip architecture

Classic server CPUs (e.g. Intel Xeon)

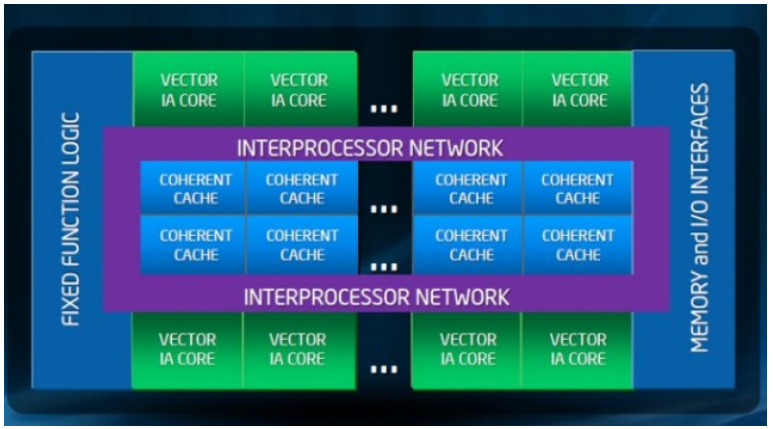


Accelerators

Heterogeneous node architecture

Accelerators

NVIDIA Kepler GPU



Intel Xeon Phi

Degree of customization

TECHNISCHE UNIVERSITÄT DARMSTADT

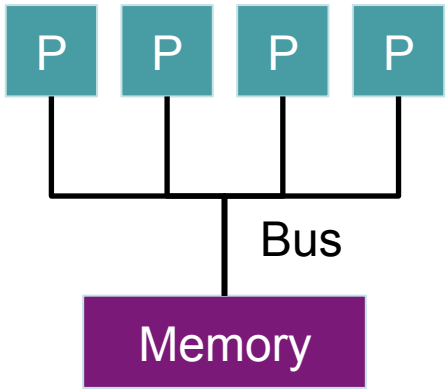
- Commodity clusters – standard nodes and standard network
  - Focus on applications with small communication requirements
  - Example: Beowulf cluster
- Custom clusters – custom nodes and custom network
  - Also called massively parallel processors
  - Focus on applications that exploit large amounts of parallelism on single problem
  - Example: IBM Blue Gene/Q
- Above classes are extremes of a broad spectrum

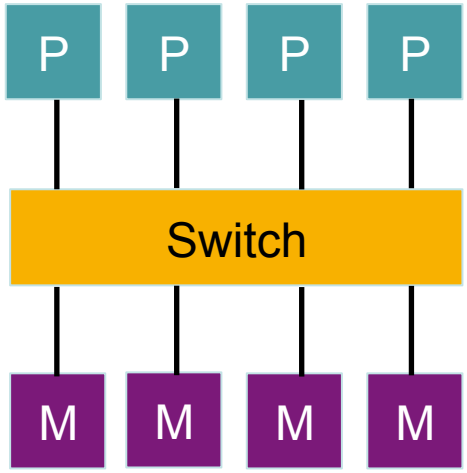
Shared memory

TECHNISCHE UNIVERSITÄT DARMSTADT

UMA (Uniform memory access)

- Each CPU has same access time to each memory address
- Simple design but limited scalability (multicore or less)



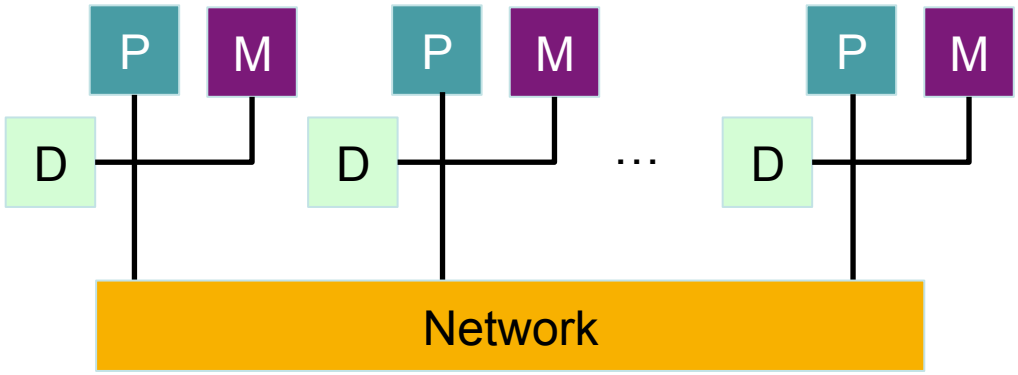


Shared memory (2)

TECHNISCHE UNIVERSITÄT DARMSTADT

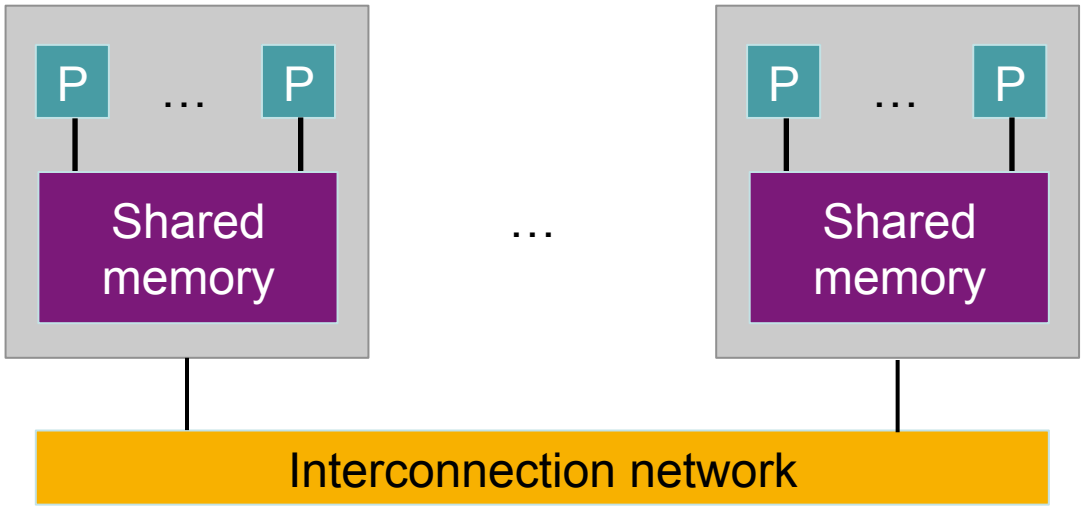
NUMA (Non-uniform memory access)

- Memory has affinity to a processor
- Access to local memory faster than to remote memory
- Harder to program but more scalable

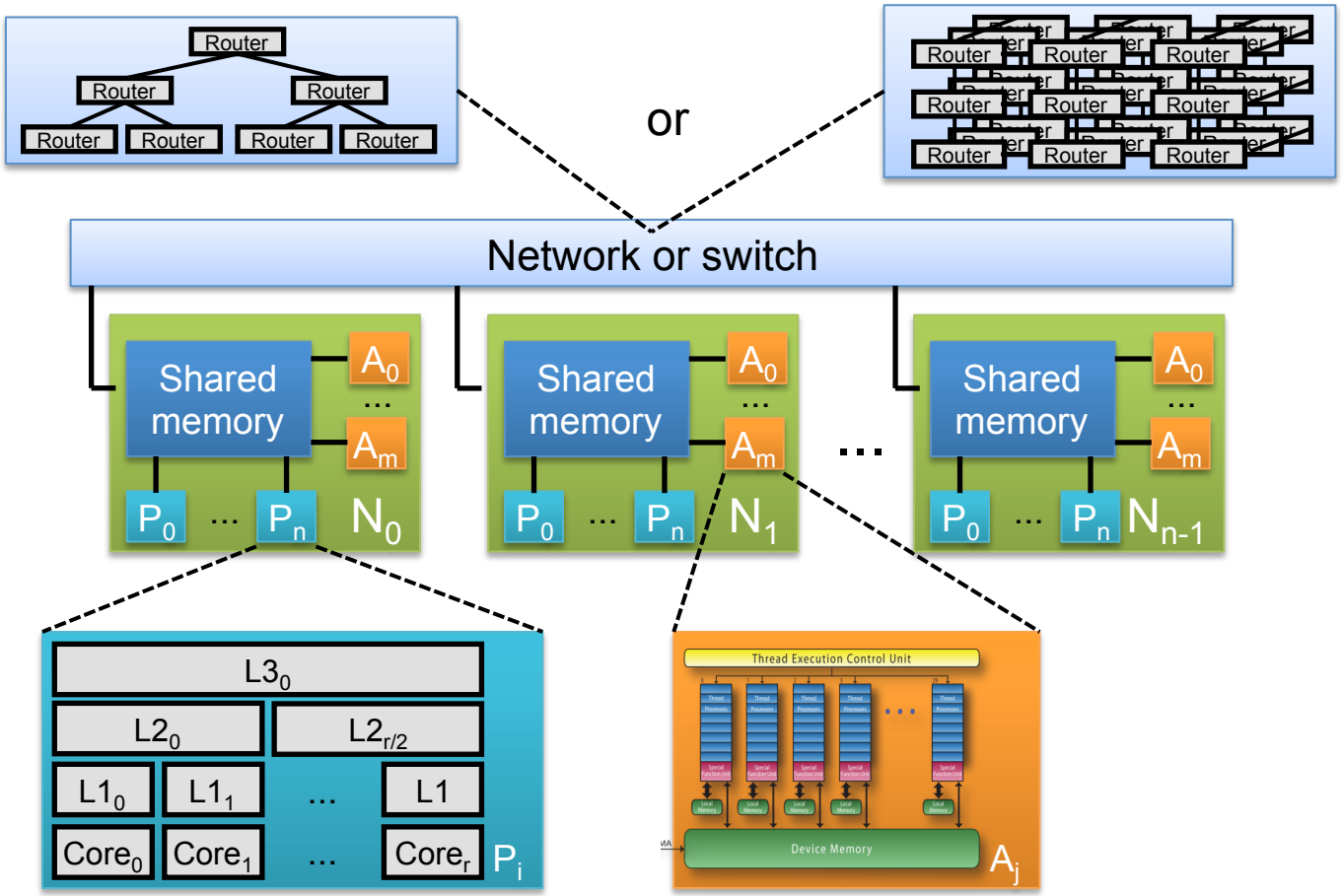


Distributed memory (aka multicomputer)

TECHNISCHE UNIVERSITÄT DARMSTADT



Typical supercomputer architecture



Interconnection network

TECHNISCHE UNIVERSITÄT DARMSTADT

Physical link between components of a parallel system

- Between processors and memory
- Between nodes

Communication via exchange of messages

- Example: intermediate results, memory requests

Design elements

- Topology – determines geometric layout of links and switches
- Routing technique – determines paths of messages through network

Network performance

TECHNISCHE UNIVERSITÄT DARMSTADT

Bandwidth

- Maximum rate at which information can be transferred
- Aggregate bandwidth – total data bandwidth supplied by network
- Effective bandwidth or throughput – fraction of aggregate bandwidth delivered to an application

Latency

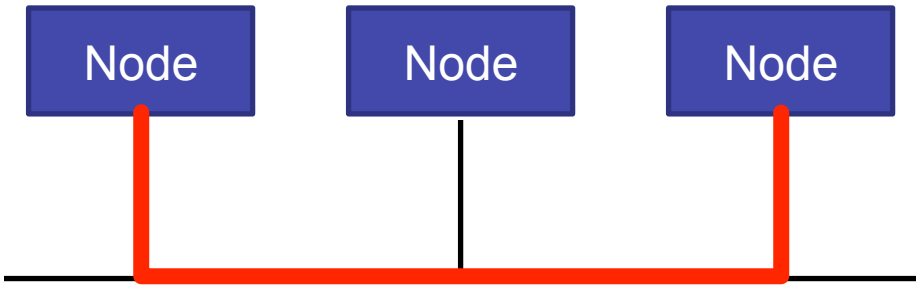
- Sending overhead + time of flight + $\frac{\text{Packet size}}{\text{Bandwidth}}$ + receiving overhead

Time for the first bit of the packet to arrive

Shared-media networks

TECHNISCHE UNIVERSITÄT DARMSTADT

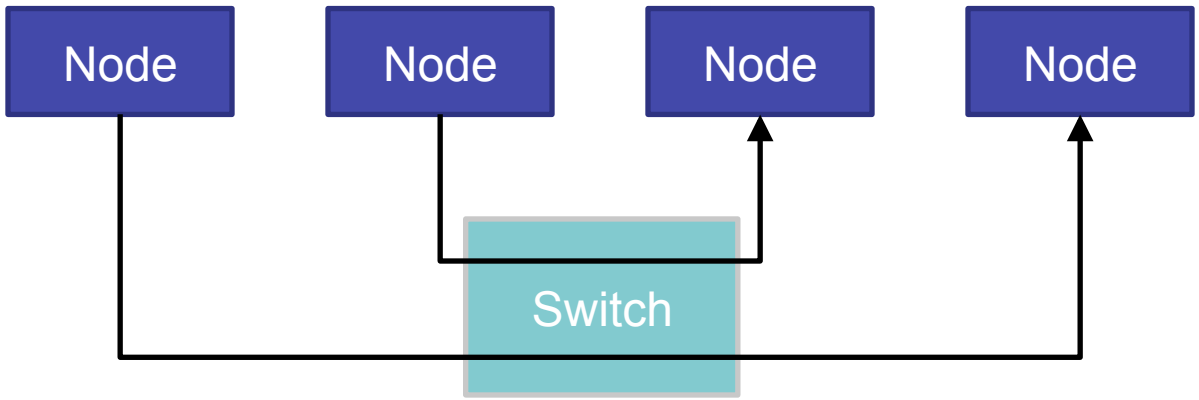
- Only one message at a time – processors broadcast their message over the medium
- Each processor “listens” to every message and receives the ones for which it is the destination
- Decentralized arbitration
  - Before sending a message, processors listen until medium is free
  - Message collision can degrade performance
- Low cost but not scalable
- Example – bus networks to connect processors to memory



Switched-media networks

TECHNISCHE UNIVERSITÄT DARMSTADT

- Support point-to-point messages between nodes
- Each node has its own communication path to the switch
- Advantages
  - Support concurrent transmission of multiple messages among different node pairs
  - Scale to very large numbers of nodes



Centralized switched networks

- Also called indirect or dynamic interconnection networks
- Connect processors / memory indirectly using several links and intermediate switches
- Examples: switching networks
- Used both for shared- and distributed-memory architectures

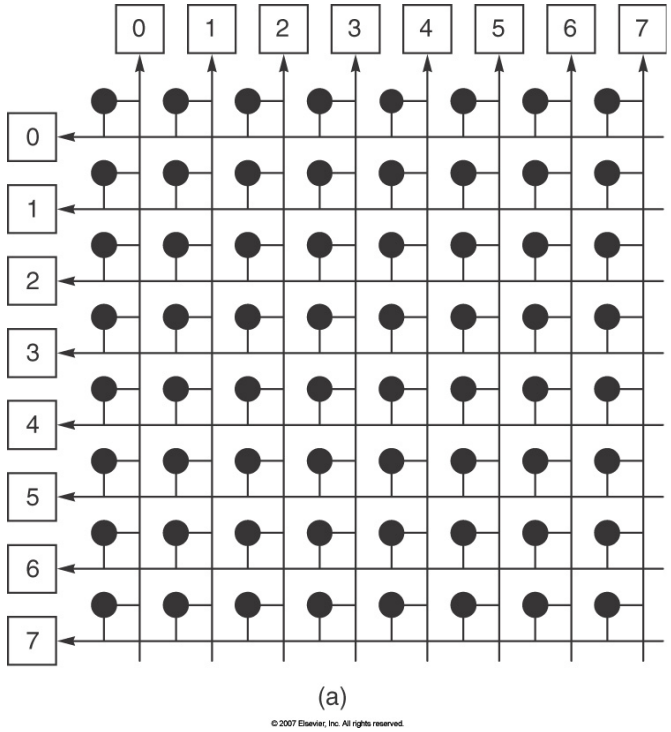
Crossbar switch

Non-blocking

- Links are not shared among paths to unique destinations

Requires N2 crosspoint switches

- Limited scalability



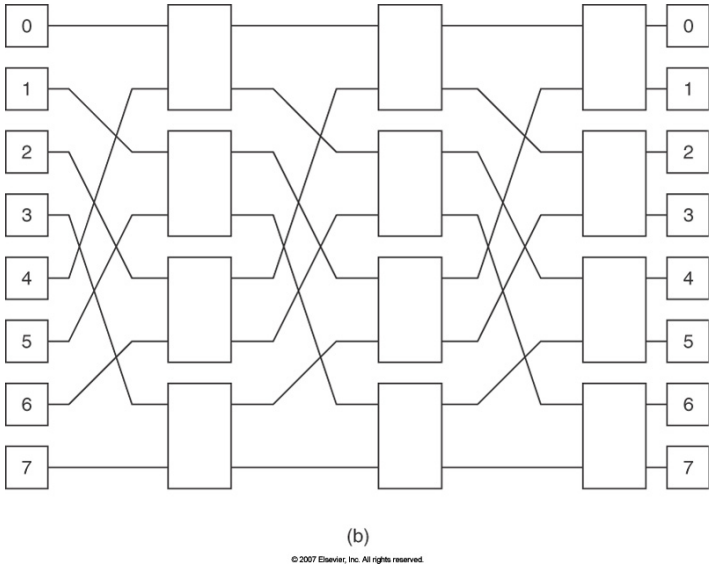
Source: Hennessy, Patterson: Computer Architecture, 4th edition, Morgan Kaufmann

Multistage interconnection network (MIN)

TECHNISCHE UNIVERSITÄT DARMSTADT

Example: Omega network

- Complexity O(N log N)
- Perfect shuffle permutation at each stage
- Blocking due to paths between different sources and destinations simultaneously sharing network links
- Omega with k x k switches
  - $\log_k N$ stages ; $N/k \log_k N$ switches



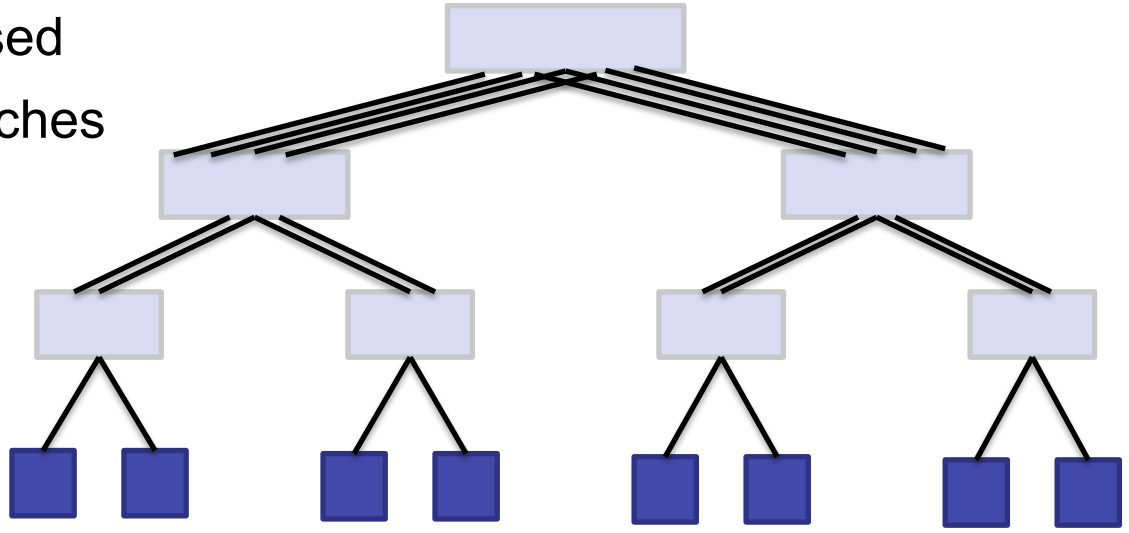
Source: Hennessy, Patterson: Computer Architecture, 4th edition, Morgan Kaufmann

MINs can be extended to rearrangeably non-blocking topologies

Fat tree

TECHNISCHE UNIVERSITÄT DARMSTADT

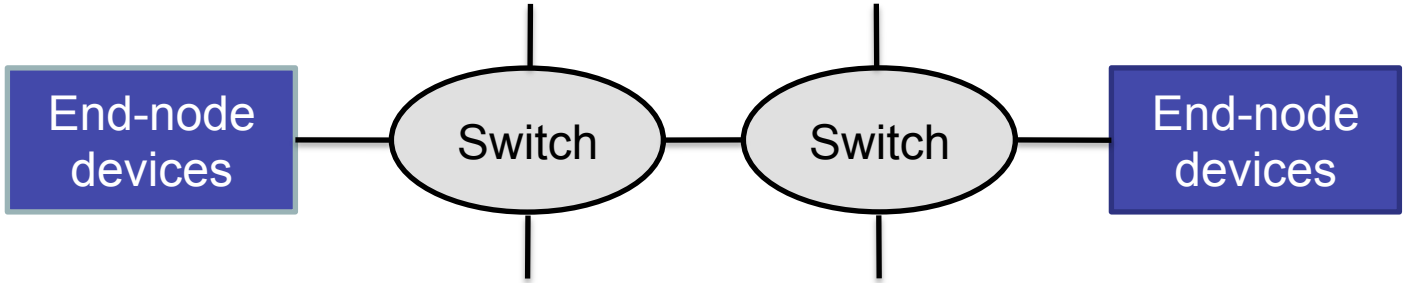
- Balanced tree where
  - Leaves = end node devices
  - Vertices = switches
- Total link bandwidth constant across all levels
- Switches often composed of multiple smaller switches
- Popular topology for cluster interconnects



Distributed switched networks

TECHNISCHE UNIVERSITÄT DARMSTADT

- Each network switch has one or more end node devices directly attached to it
- End node devices = processor(s) + memory
  - Directly connected to other nodes without going through external switches
  - Mostly used for distributed-memory architectures
- Also called direct or static interconnection networks
- Ratio of switches to nodes = 1:1



Evaluation criteria

TECHNISCHE UNIVERSITÄT DARMSTADT

Network degree

- Maximum node degree
- Node degree = number of adjacent nodes = (incoming + outgoing) edges

Diameter

- Largest distance between two nodes

Bisection width

- Minimum number of edges between nodes that must be removed to cut the network into two roughly equal halves
- Bisection bandwidth = bandwidth [bytes/s] between the two parts

Edge / node connectivity

- Minimum number of edges / nodes that need to be removed to render network disconnected

Embedding

- Mapping of one network onto another

Requirements

TECHNISCHE UNIVERSITÄT DARMSTADT

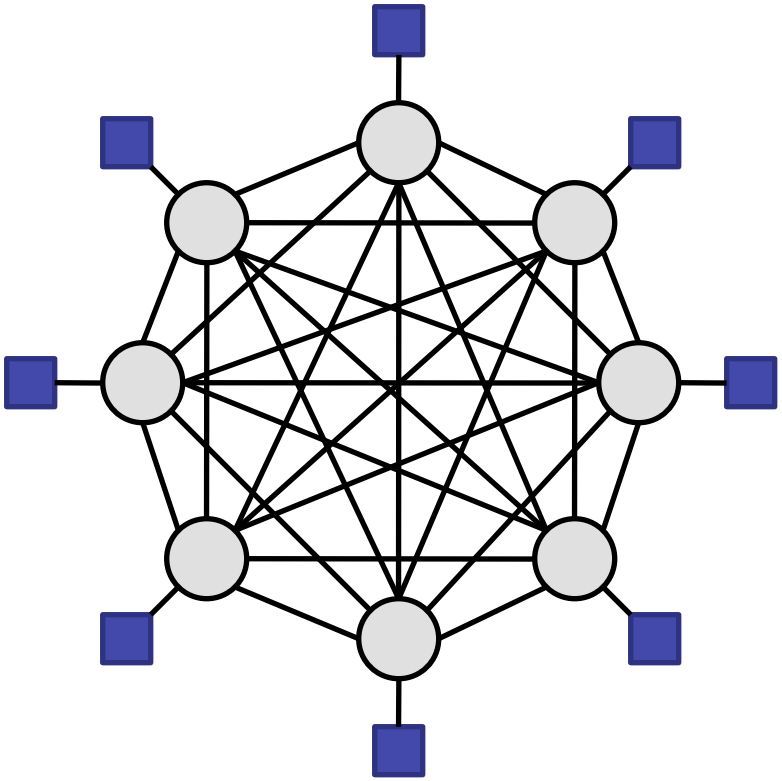
- Low network degree to reduce hardware costs
- Low diameter to ensure low distance (i.e., latency) for message transfer
- High bisection bandwidth to ensure high throughput
- High connectivity to ensure robustness
- Option to embed many other networks to ensure flexibility

Often conflicting goals

Fully connected topology

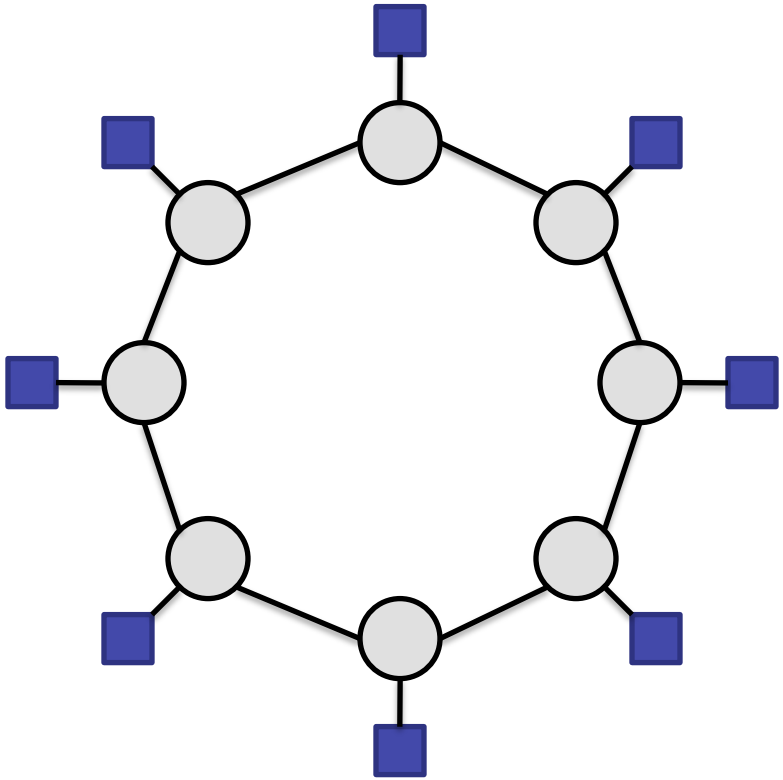
TECHNISCHE UNIVERSITÄT DARMSTADT

- Each node is directly connected to every other node
- Expensive for large numbers of nodes
- Dedicated link between each pair of nodes
- Cheaper alternative: crossbar topology



Ring topology

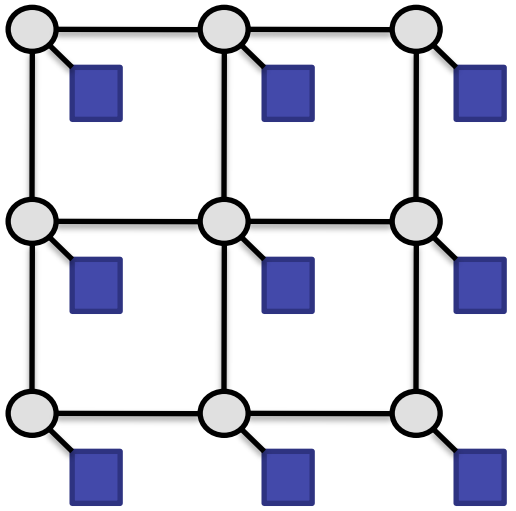
TECHNISCHE UNIVERSITÄT DARMSTADT



N-dimensional meshes

TECHNISCHE UNIVERSITÄT DARMSTADT

- Typically 2 or 3 dimensions
- Direct link to neighbors
- Each node has 1 or 2 neighbors per dimension
  - 2 in the center
  - Less for border or corner nodes
- Efficient nearest neighbor communication
- Suitable for large numbers of nodes

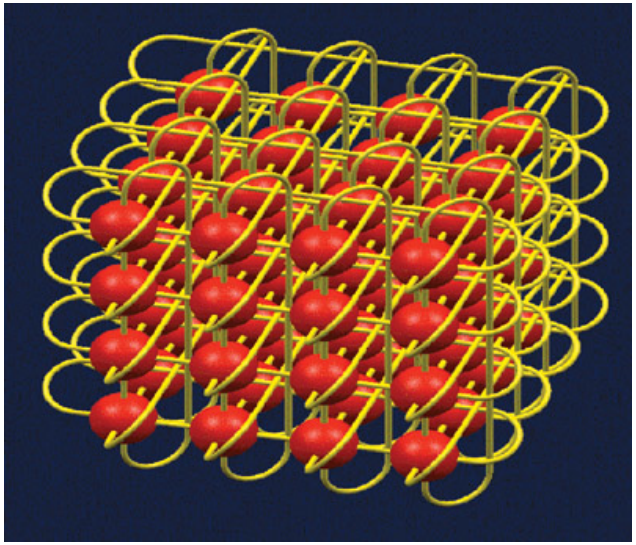


2D mesh

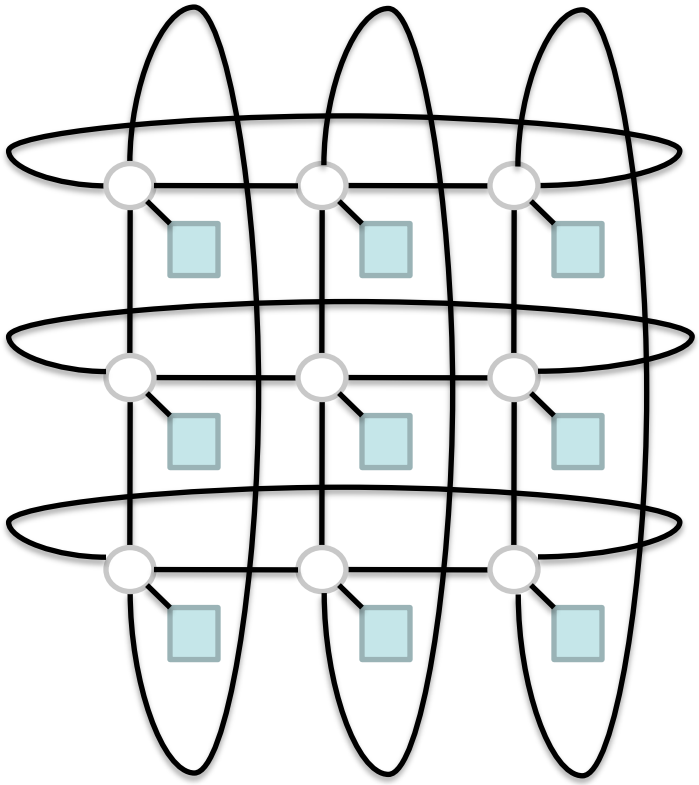
Torus

TECHNISCHE UNIVERSITÄT DARMSTADT

- Mesh with wrap-around connections
- Each node has exactly 2 neighbors per dimension



3D torus

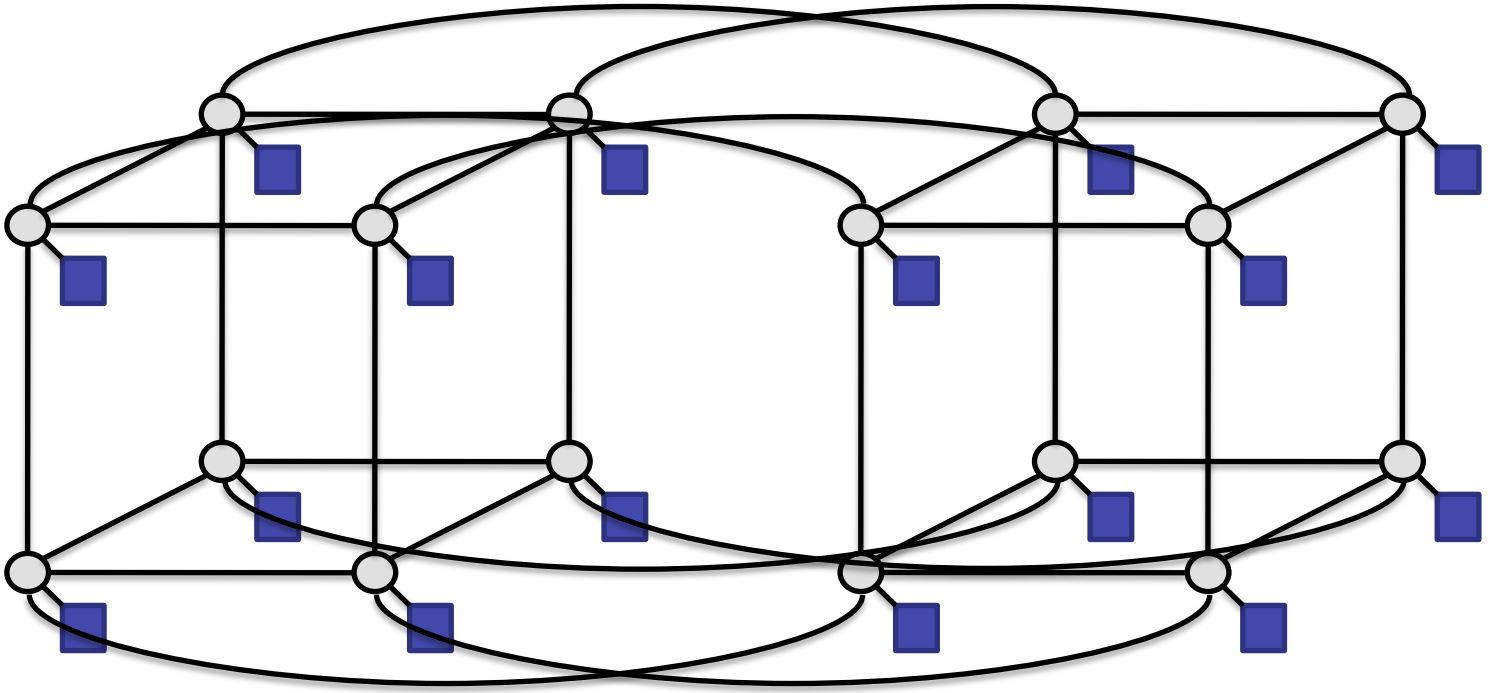


2D torus

Hypercube

16 nodes
(16 = 2^4 so n = 4)

TECHNISCHE UNIVERSITÄT DARMSTADT



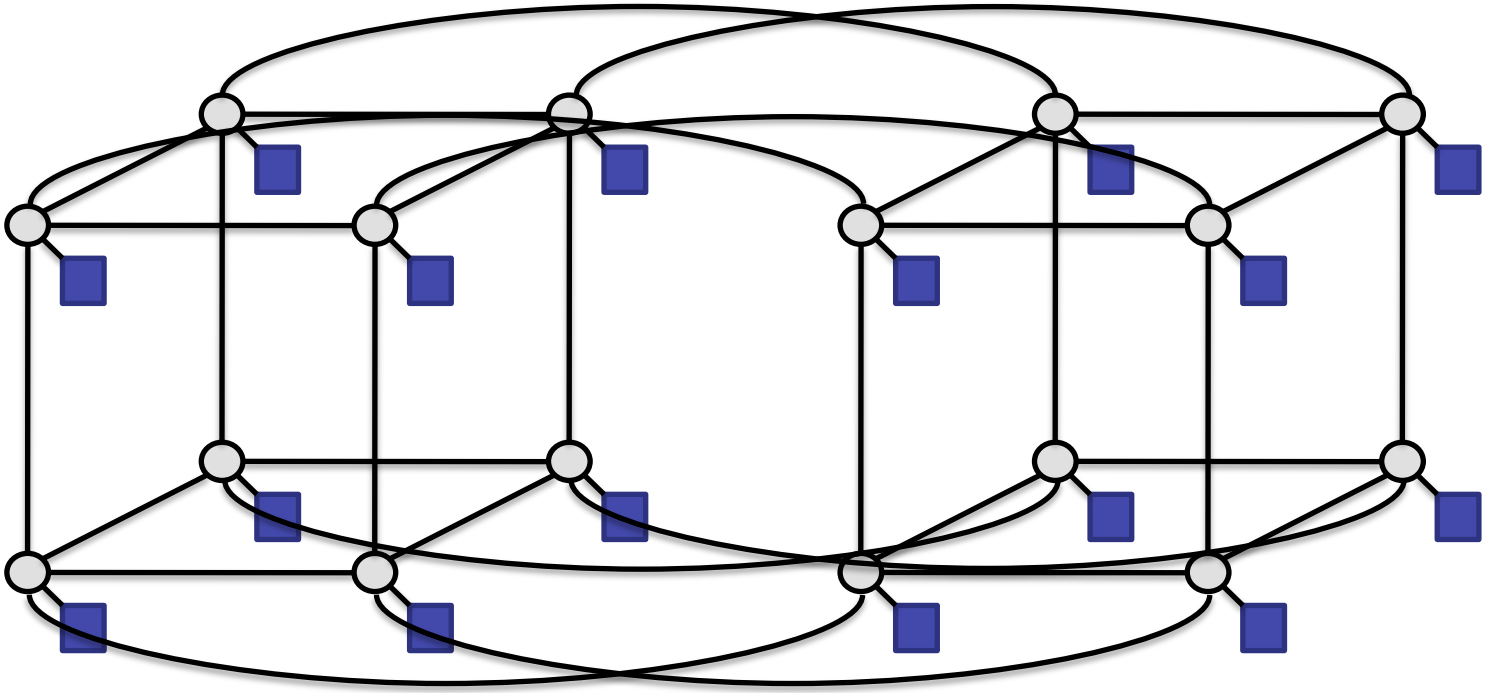
Each node has one connection along each dimension (n = #dimensions)

Usually better connectivity than tori at the expense of higher link and switch costs

Hypercube

16 nodes
(16 = 2^4 so n = 4)

TECHNISCHE UNIVERSITÄT DARMSTADT



Each node has one connection along each dimension (n = #dimensions)

Usually better connectivity than tori at the expense of higher link and switch costs

Performance and cost (64 nodes)

TECHNISCHE UNIVERSITÄT DARMSTADT

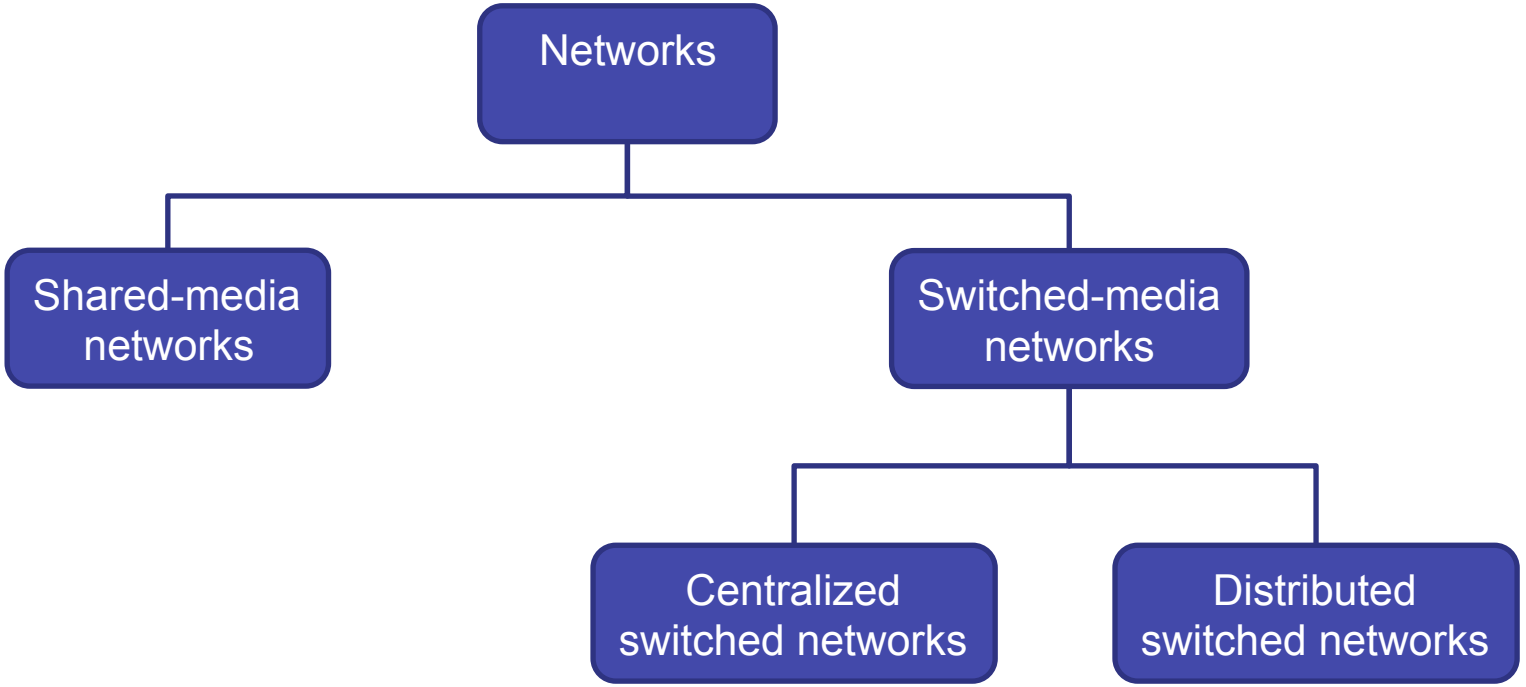
Evaluation category	Bus	Ring	2D mesh	2D torus	Hypercube	Fat tree	Fully connected
Performance							
BW bisection [#links]	1	2	8	16	32	32	1024
Max (avg.) hop count	1(1)	32(16)	14(7)	8(4)	6(3)	11(9)	1(1)
Cost							
I/O ports per switch	NA	3	5	5	7	4	64
#Switches	NA	64	64	64	64	192	64
#Network links	1	64	112	128	192	320	2016
Total #links	1	128	176	192	256	384	2080

Commercial HPC machines

TECHNISCHE UNIVERSITÄT DARMSTADT

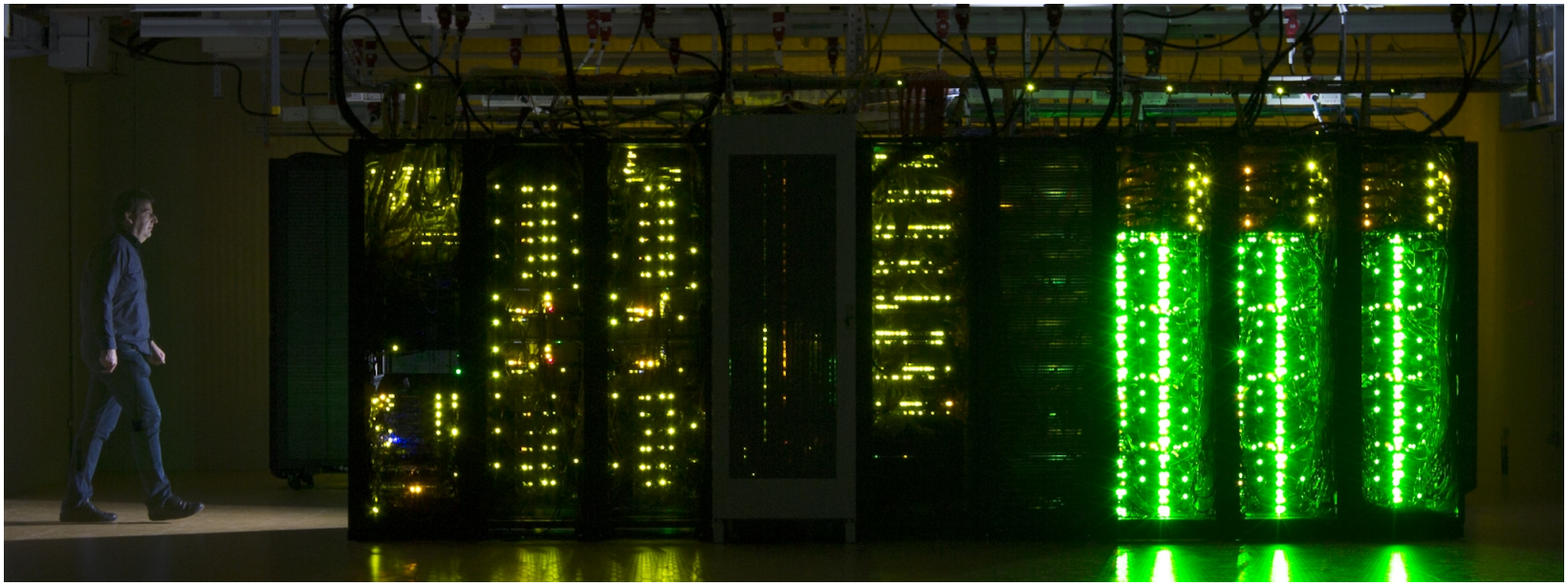
Company	System name [network name]	Max. #nodes [x #CPUs]	Basic network topology
Intel	ASCI Red Paragon	4816 [x 2]	2D mesh 64 x 64
IBM	ASCI White SP Power3 [Colony]	512 [x 16]	Bidirectional MIN (fat tree or Omega)
Intel	Thunder Itanium2 Tiger 4 [QsNet^{II}]	1024 [x 4]	Fat tree with 8-port bi-directional switches
Cray	XT3 [Seastar]	30,508 [x 1]	3D torus 40 x 32 x 24
Cray	X1E	1024 [x 1]	4-way bristled 3D torus (~23 x 11)
IBM	ASC Purple pSeries 575 [Federation]	> 1280 [x 8]	Bidirectional MIN (fat tree or Omega)
IBM	Blue Gene/L eServer Solution [Torus Network]	65,536 [x 2]	3D torus 32 x 32 x 64

Network classes



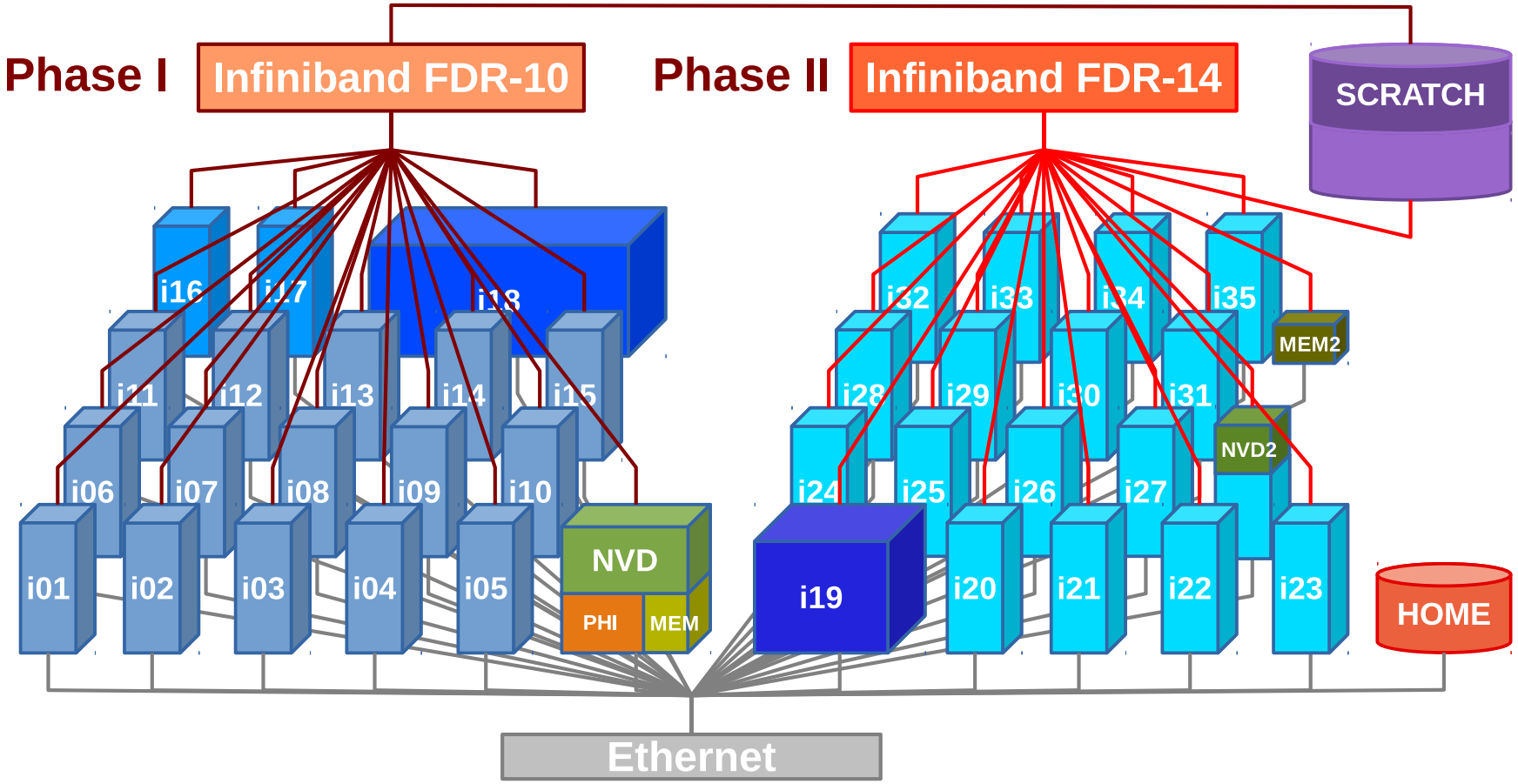
Lichtenberg Cluster @ TU Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT



One cluster – multiple islands

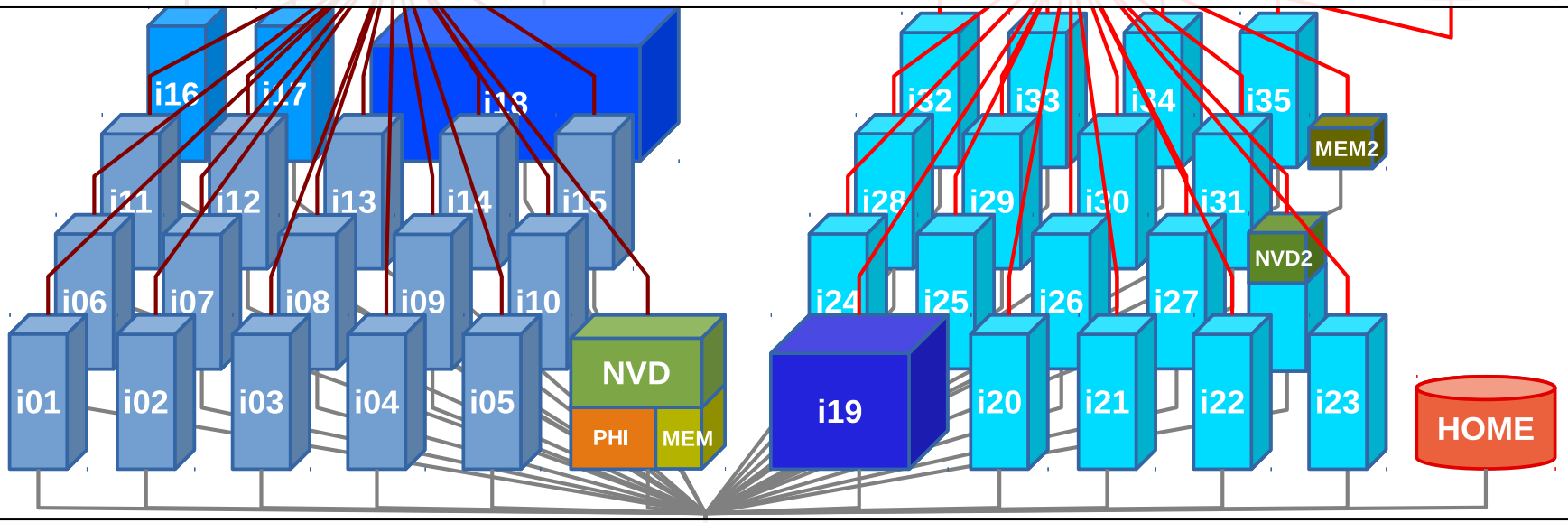
TECHNISCHE UNIVERSITÄT DARMSTADT



One cluster – multiple islands

TECHNISCHE UNIVERSITÄT DARMSTADT

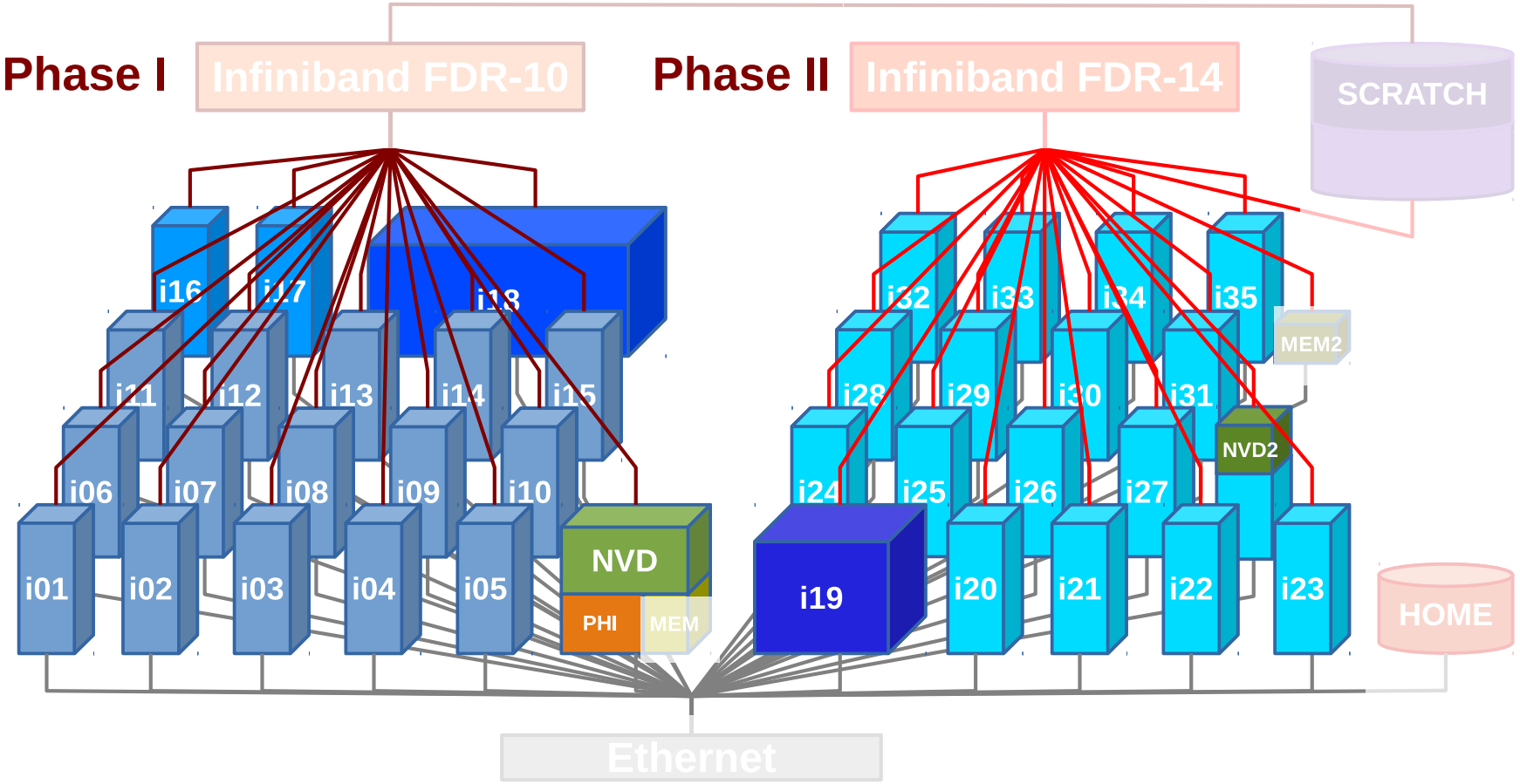
- Cluster is divided into 2 phases
- Each phase is divided into several islands
- Rule of thumb:
  1 island ≙ 32 compute nodes ≙ 512 (ph. I) / 768 (ph. 2) CPU cores
- For large computations, there are 2 islands with more than 2000 CPU cores



- Computation across more than one island is only possible on request, due to some technical limitations (across phases impossible).

Compute nodes (“mpi”, “nvd”, “phi”)

TECHNISCHE UNIVERSITÄT DARMSTADT



Compute nodes

TECHNISCHE UNIVERSITÄT DARMSTADT

Phase I (704+70 nodes):

Processors:

- 2 Intel Xeon E5-4650 (**Sandy Bridge**) processors
  ≙ 2 · 8 = **16 CPU cores**
- **2.7 GHz**
  (up to 3.3 GHz in turbo mode)

Main Memory:

- **32 GB RAM** (some have 64 GB)

Network:

- Gigabit Ethernet
- FDR-10 InfiniBand

Phase II (596+31 nodes):

Processors:

- 2 Intel Xeon E5-2680 v3 (**Haswell**) processors
  ≙ 2 · 12 = **24 CPU cores**
- **2.5 GHz**
  (up to 3.3 GHz in turbo mode)

Main Memory:

- **64 GB RAM**

Network:

- Gigabit Ethernet
- FDR-14 InfiniBand

Accelerator nodes

TECHNISCHE UNIVERSITÄT DARMSTADT

NVIDIA nodes

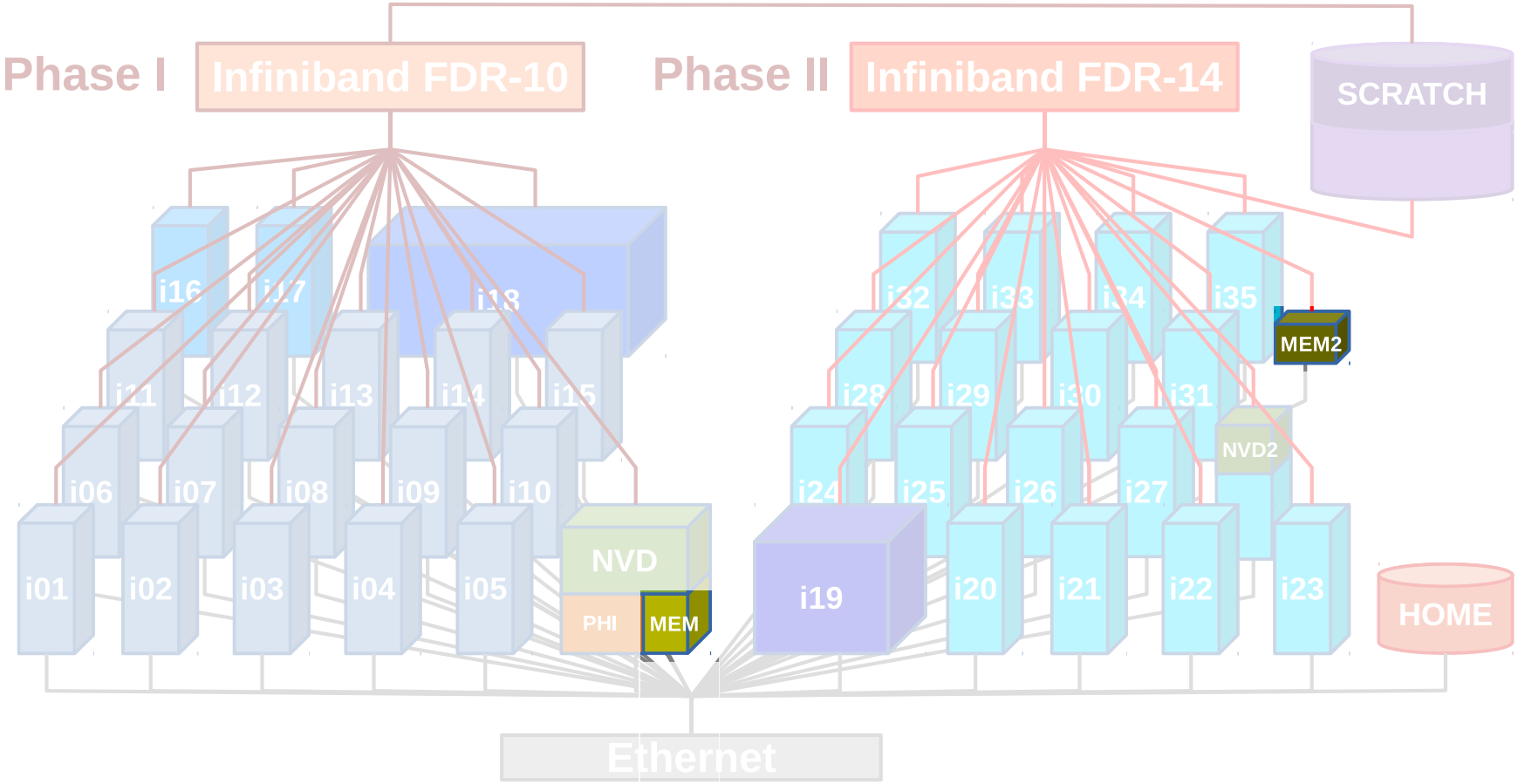
- 44 Sandy Bridge compute nodes have 2 **NVIDIA K20Xm** cards each
- 2 Haswell compute nodes have 2 **NVIDIA K40m** cards each
- 1 Haswell compute node has 2 **NVIDIA K80** cards

Xeon Phi Nodes

- 24 Sandy Bridge compute nodes have
  2 **Intel Xeon Phi** 5110P cards each
- 2 Sandy Bridge compute nodes have
  2 **Intel Xeon Phi** 7120P cards each

Big mem nodes (“mem”, “mem2”)

TECHNISCHE UNIVERSITÄT DARMSTADT



Mem nodes

TECHNISCHE UNIVERSITÄT DARMSTADT

Phase I (4 nodes):

Processors:

- 8 Intel Xeon E7-8837 (**Westmere**) processors
  ≙ 8 · 8 = **64 CPU cores**
- **2.66 GHz**
  (up to 2.8 GHz in turbo mode)

Main Memory:

- **1 TB** (1024 GB) **RAM**

Network:

- 10 Gigabit Ethernet
- 2 · FDR-10 InfiniBand

Phase II (4 nodes):

Processors:

- 4 Intel Xeon E7-4890 v2 (**Ivy Bridge**) processors
  ≙ 4 · 15 = **60 CPU cores**
- **2.8 GHz**
  (up to 3.4 GHz in turbo mode)

Main Memory:

- **1 TB** (1024 GB) **RAM**

Network:

- 10 Gigabit Ethernet
- 2 · FDR-14 InfiniBand

File systems

TECHNISCHE UNIVERSITÄT DARMSTADT

Mountpoint	/home	/work/scratch	/work/local
Size	> 300 TB	> 650 TB	> 100 GB per node
Access time	Normal (Ethernet)	Fast (InfiniBand)	Very fast (local HDD)
Accessibility	Global (cluster)	Global (cluster)	Local (node)
Data availability	permanent	≥ 1 month	Only during job runtime
Quota*	15 GB**	100 TB** 2 Mio. files**	none
Backup	Weekly + snapshots	none	none

* Use the command **`cquota`** to find out your current usage and quota.
** Can be increased on request.

Login nodes

TECHNISCHE UNIVERSITÄT DARMSTADT

4 nodes (hardware similar to Phase I):

Processors:

- 4 Intel Xeon E5-4650 (**Sandy Bridge**) processors
  ≙ 4 · 8 = **32 CPU cores**
- **2.7 GHz**
  (up to 3.3 GHz in turbo mode)

Main Memory:

- **128 GB RAM**

Network:

- 2 · 10 Gigabit Ethernet
- 2 · FDR-10 InfiniBand

8 nodes (hardware similar to Phase II):

Processors:

- 2 Intel Xeon E5-2680 v3 (**Haswell**) processors
  ≙ 2 · 12 = **24 CPU cores**
- **2.5 GHz**
  (up to 3.3 GHz in turbo mode)

Main Memory:

- **128 GB RAM**

Network:

- 2 · 10 Gigabit Ethernet
- FDR-14 InfiniBand

IBM BlueGene/Q

TECHNISCHE UNIVERSITÄT DARMSTADT

Blue Gene/Q JUQUEEN at Forschungszentrum Jülich

Blue Gene design goals

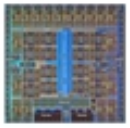
TECHNISCHE UNIVERSITÄT DARMSTADT

- System-on-Chip (SoC) design
- Processor comprises both processing cores and network
- Optimal performance / watt ratio
- Small foot print
- Transparent high-speed reliable network
- Easy programming based on standard message passing interface (MPI)
- Extreme scalability (> 1.5 M cores)
- High reliability

Blue Gene/Q design

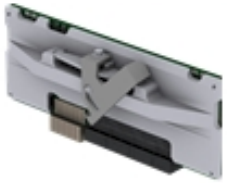
TECHNISCHE UNIVERSITÄT DARMSTADT

1. Chip
16 cores

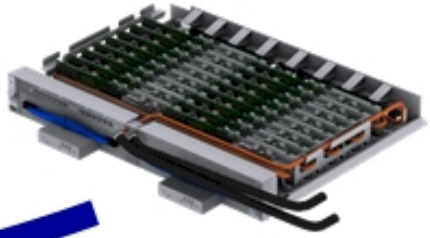


2. Module
Single Chip

3. Compute Card ("Node"):
One BQC Module (1x1x1x1x1),
16 GB DDR3 Memory



4. Node Card ("Node Board"):
32 Compute Cards (2x2x2x2x2),
Optical Modules, BQL Link Chips, Torus



5a. Midplane:
16 Node Cards (4x4x4x4x2)

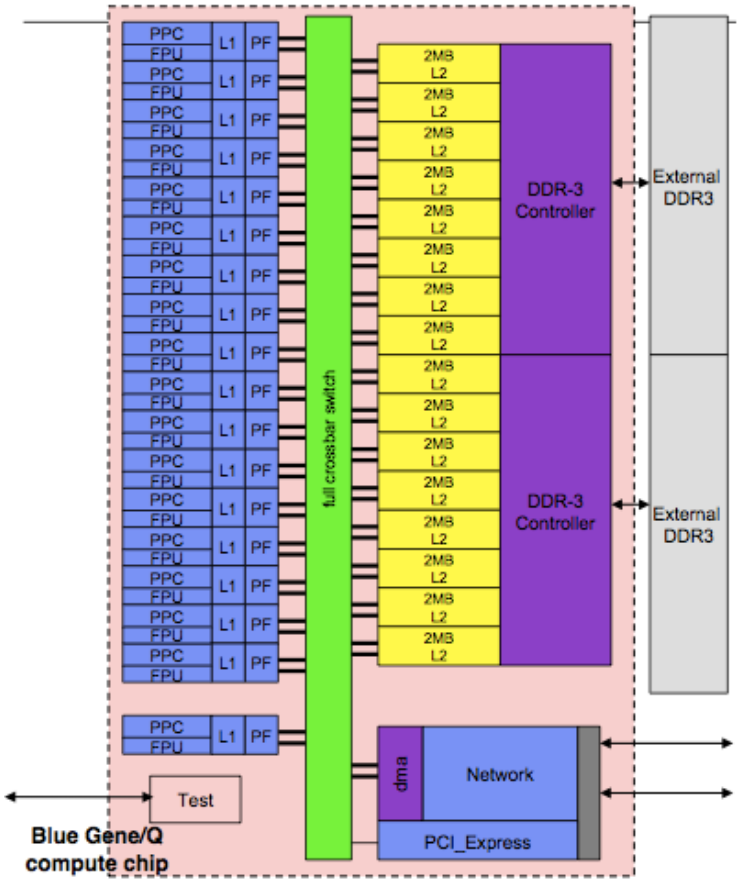
5b. I/O drawer (1,2 or 4 per rack):
* 8 I/O cards @ 16 GB,
* 8 PCIe gen2 x8 slots (IB, 10GbE)

6. Rack: 2 Midplanes (4x4x4x8x2)
1,2 or 4 I/O Drawer

7. System:
e.g. 8 racks (8x8x8x8x2) = 1.7 PF/s
e.g. 28 racks (8x28x8x8x2) = 5.9 PF/s
e.g. 96 racks (16x12x16x16x2) = 20 PF/s

Source: IBM

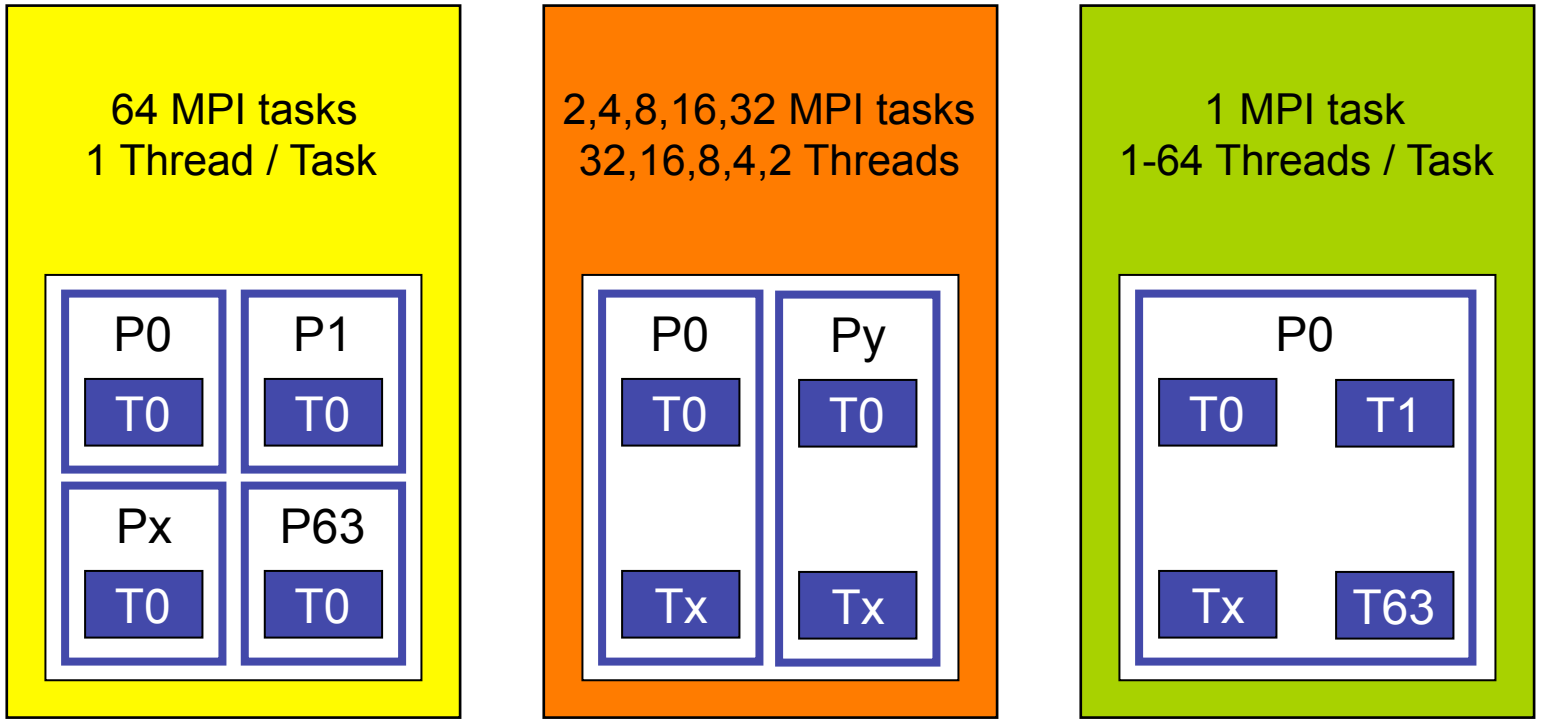
Blue Gene/Q chip architecture



- 16+1+1 core SMP @ 1.6 GHz
  - Each core 4-way hardware threaded
  - 2-way concurrent issue
- Transactional memory and thread level speculation
- Quad floating point unit on each core
  - 204.8 GF peak per node
- 563 GB/s bisection bandwidth to shared L2
- 32 MB shared L2 cache
- 42.6 GB/s DDR3 bandwidth (1.333 GHz DDR3)
- 10 intra-rack inter-processor links each at 2.0GB/s (5D-Torus)
- One I/O link at 2.0 GB/s
- 16 GB memory/node
- ~60 watts max chip power

Execution Modes in BG/Q

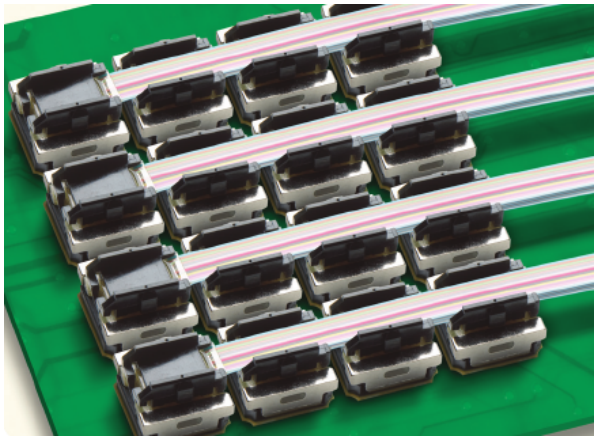
TECHNISCHE UNIVERSITÄT DARMSTADT



BG/Q's new network architecture

TECHNISCHE UNIVERSITÄT DARMSTADT

- 11 bi-directional chip-to-chip links
  - 2 GB/s bandwidth, about 40 ns latency
- 5-dimensional torus topology
  - Dimension E limited to length 2
- Why d-dimensional torus with large d?
  - High bi-section bandwidth
  - Flexible partitioning in lower dimensions
- Deterministic/dynamic routing support
- Collective and barrier networks embedded in 5-D torus network
  - Floating point addition support in collective network
  - 11th port for auto-routing to IO fabric



Source: IBM

JUQUEEN Configuration

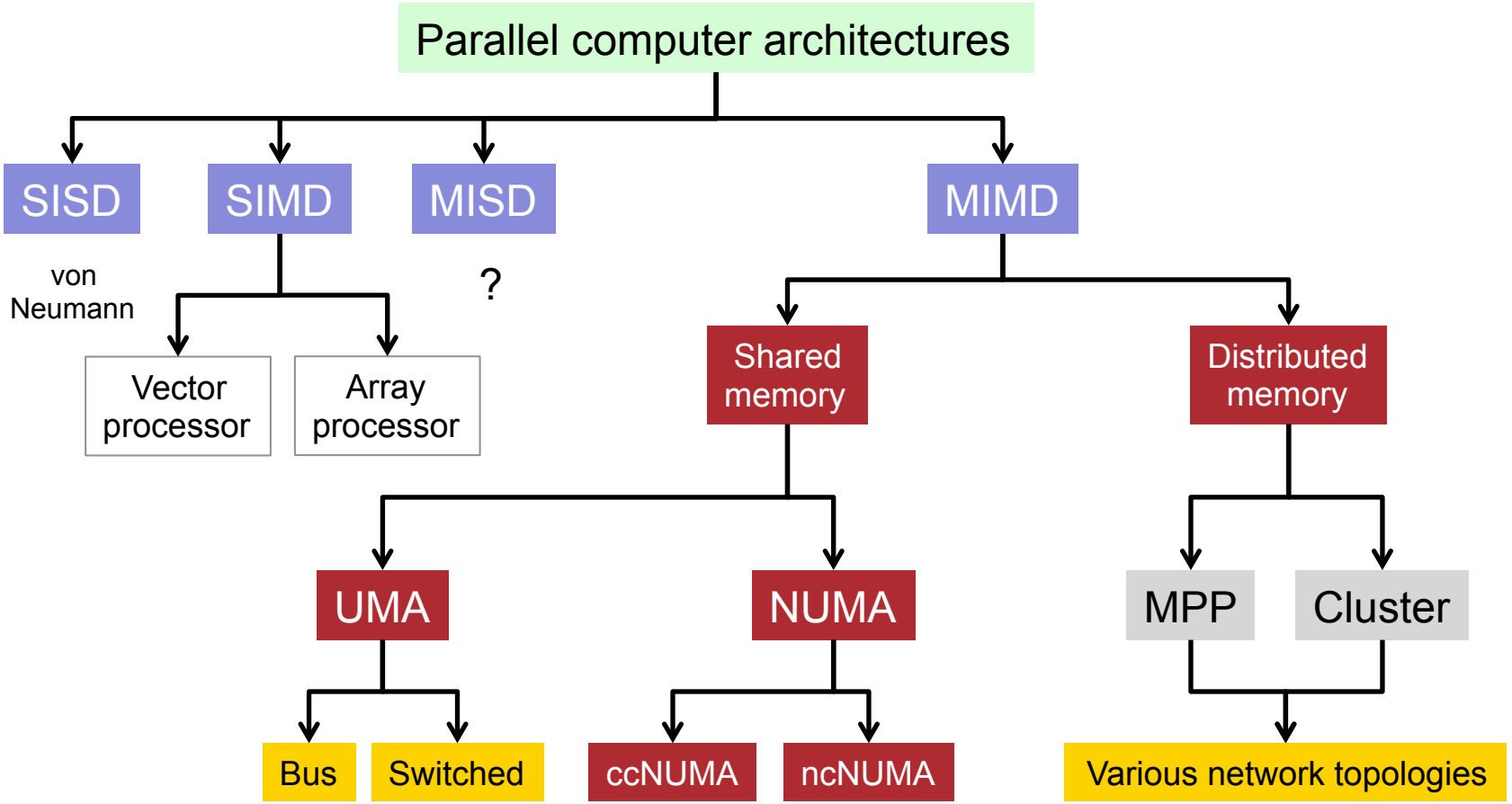
28 racks Blue Gene/Q

- 28672 compute nodes (16 cores, 16 GB memory)
- 458752 cores / 1.8M threads
- 5.88 PFlop/s peak performance
- 248 I/O nodes (10GigE) ← (1x32 + 27x8)
- ~2.2 MW power consumption (~80 kW per rack)

4 frontend nodes (user login) + 2 service nodes (system, database)

- IBM p7 740, 8 cores (3.55 GHz), 128 GB memory
- Local storage device DS5020 (16 TB)

Summary


Parallel computer architectures
SISD
SIMD
MISD
MIMD
von Neumann
?
Vector processor
Array processor
Shared memory
Distributed memory
UMA
NUMA
MPP
Cluster
Bus
Switched
ccNUMA
ncNUMA
Various network topologies