HUST-USYD Summer School on Parallel Programming Practice – Lecture 1

Bing Bing Zhou (bing.zhou@sydney.edu.au)

School of Computer Science, University of Sydney

THE UNIVERSITY OF SYDNEY

The Course

- Lecturers: introduce basic methods for parallel algorithm design and implementation
- Students: design and implement parallel algorithms to solve several simple problems
- Objectives:
  - have a good understanding of basic concepts in parallel computing
  - Lay a solid foundation for further in-depth learning of parallel computing theory and its application in the future

Lecturers and Tutors

- Lecturers:
  - Bing Bing Zhou (周兵兵）, bing.zhou@sydney.edu.au
  - Feng Lu (陆枫，also Coordinator), lufeng@hust.edu.cn
- Tutors:
  - Hao Han（韩浩）, hanhao@hust.edu.cn
  - Yuxiang Hou（侯宇翔）， houyx@hust.edu.cn

Outline

- Parallel computing
  - Definition
  - Why parallel computing
- Parallel computer organization
- Interconnection networks
  - Frontier - the world’s first exascale supercomputer
- Parallel computer classification
- Parallel programming models
- Lab exercise: getting familiar with EduCoder (头歌平台)

Parallel Computing

- **Simple definition:** Using parallel computers, computer clusters, or other advanced parallel/distributed computing systems to solve advanced computation problems at high speed
  - The performance of parallel computers is commonly measured in floating-point operations per second (FLOPS)
  - E.g., giga (10^9) FLOPS, tera (10^{12}) FLOPS, peta (10^{15}) FLOPS, exa (10^{18}) FLOPS
  - Theoretical peak FLOPS on a single CPU can be calculated
    - $FLOPS = cores \times \frac{cycles}{second} \times \frac{floating{-}point\ operations}{cycle}$

Parallel Computing

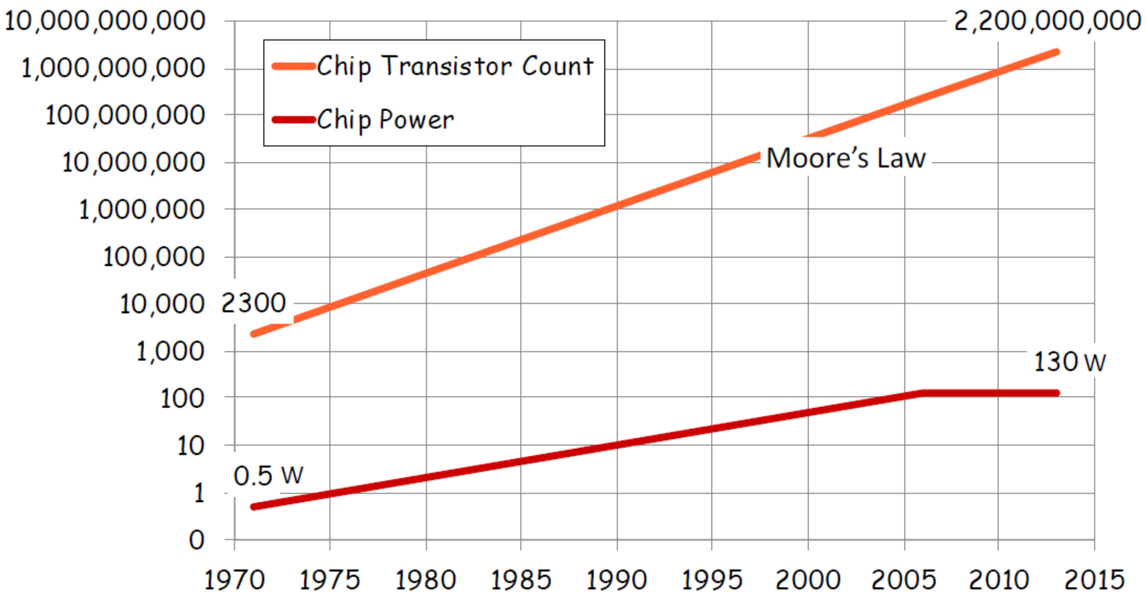
- Parallel computing works by dividing a large task into multiple smaller ones and then assign each small task to a processor
  - Note during the computation these small tasks may need coordination, making parallel computing more complicated
- Problems to consider:
  - How to divide a task into multiple smaller ones?
  - How do we assign tasks to processes/threads?
  - What if processes/threads need to share partial results?
  - What is the performance of a parallel program?
  - …

Parallel Computing

- Why parallel computing?
  - Technology push
    - Nowadays even our laptops, or mobile phones are parallel computers – every computer is a parallel computer
  - Application driven
    - Many modern applications, such as big data analytics, artificial intelligence and deep learning, heavily rely on parallel computing

Technological Limitations

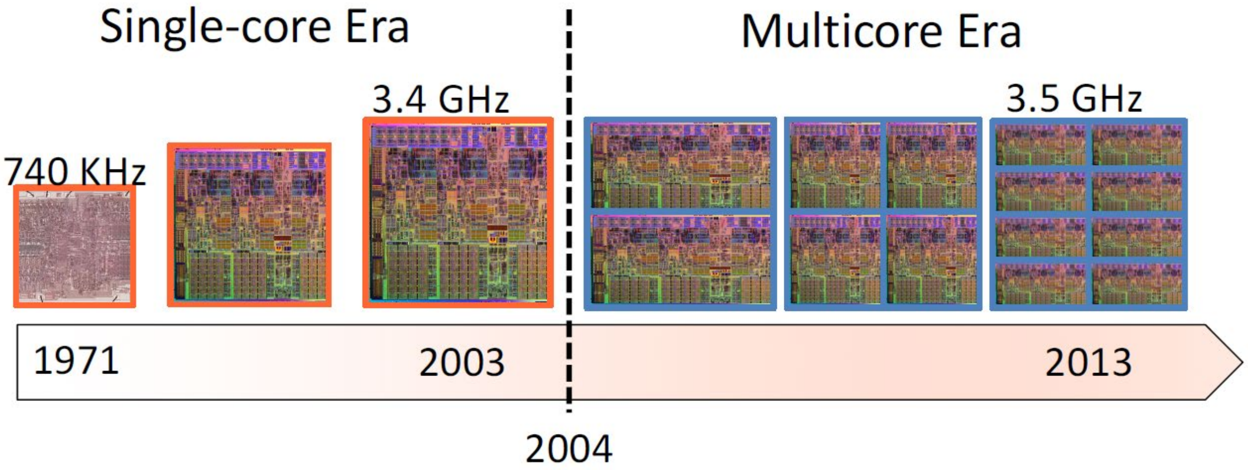
- "Frequency wall": Increasing frequencies and deeper pipelines has reached diminishing returns on performance
  - "Power wall": The chip will melt if running any faster (higher clock rate)
  - "ILP wall": There are diminishing returns on finding more ILP (instruction-level parallelism)
- Moore's law is alive and well
- However, cannot significantly increase frequency to increase performance



Powering the transistors without melting the chip

Multicore Technology

- $power \propto voltage^2 \times frequency\ (v^2 f)$
- $frequency \propto voltage$
- $power \propto frequency^3$
- For **single core**, increase frequency by 50%, we will have 1.5x performance, but 3.3x power
- Using **two cores**, to increase same peak performance we may actually decrease frequency by 25%, and then reduce power to 0.8x



Computing Applications

- When computing cost improves, the opportunities of computers multiply
- Science
  - storm forecasting and climate prediction
  - understanding biochemical processes of living organisms
  - …
- Engineering
  - computational fluid dynamics and airplane design
  - earthquake and structural modelling
  - molecular nanotechnology
  - …
- Business
  - computational finance
  - …

Big Data (Data Analytics)

- Data deluge is commonly seen nowadays in many domains, such as astronomy, particle physics, smart city and e-health
- The dramatic increase in the volume of data also requires a large amount of computing power necessary to transform the raw data into meaningful information, which demands the use of parallel and distributed computing strategies
- These demanding requirements have led to the development of high-level programming models such as MapReduce to make the parallelization of data-intensive computation over many machines simple

Deep Learning

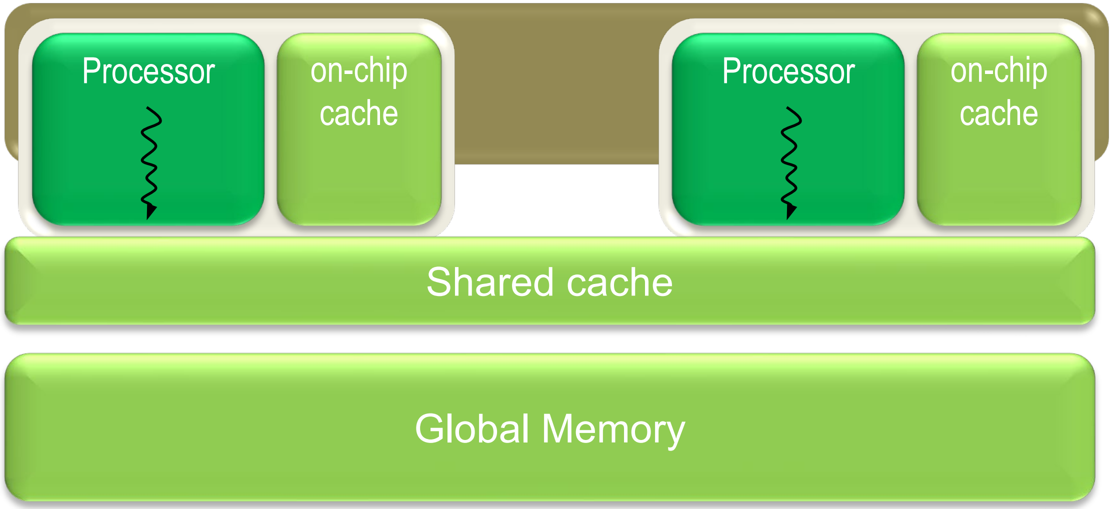
- Deep learning (DL), or Deep Neural Network (DNN) is a class of machine learning algorithms which are inspired by the structure and function of the human brains
- DNNs are trained by analysing a large amount of data to enable the classification and prediction
- It has now been successfully applied for various applications in practice and play more and more significant roles in our daily lives
- However, the great success is mainly due to the increased amount of available training data and more powerful computing resources which enable us to train larger and deeper neural networks

Parallel Computer Organization

- Three types of physical organization:
  - Multicore
    - Multiple cores with a shared memory
    - Small scale (~100 cores at high end)
  - Cluster
    - Multiple processing elements (PEs), stand-alone computers with multicore and/or GPUs, connected via interconnection networks
    - Large scale, almost all supercomputers are clusters
  - GPU
    - Attached accelerator - originally designed for graphics processing
    - Nowadays can be used as a kind of more general-purpose GPU for applications with regular computation and data pattens, especially for machine learning

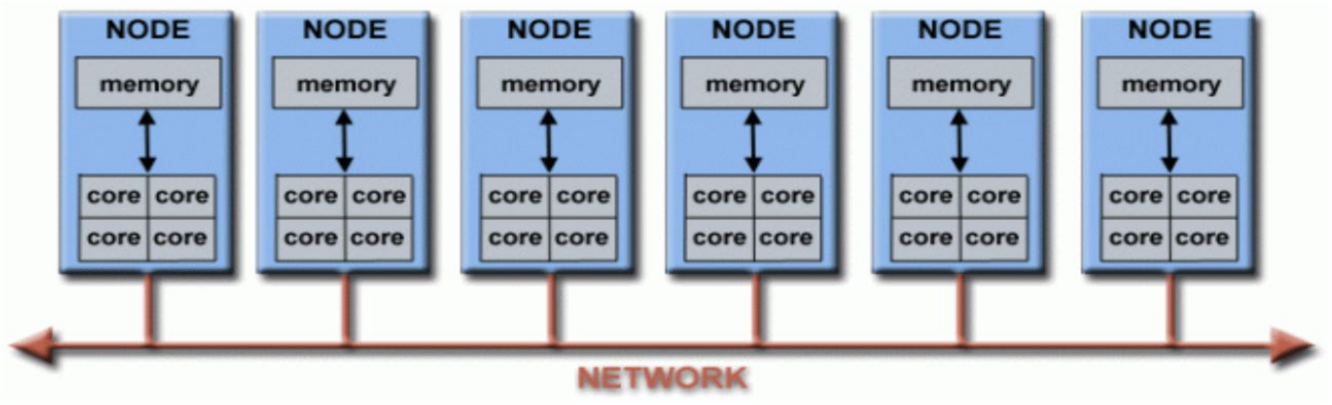
Multicore

- All computers nowadays are multicore computer system:
  - A number of processors, or cores
  - On-chip cache
  - Shared global memory space (external cache and DRAM)



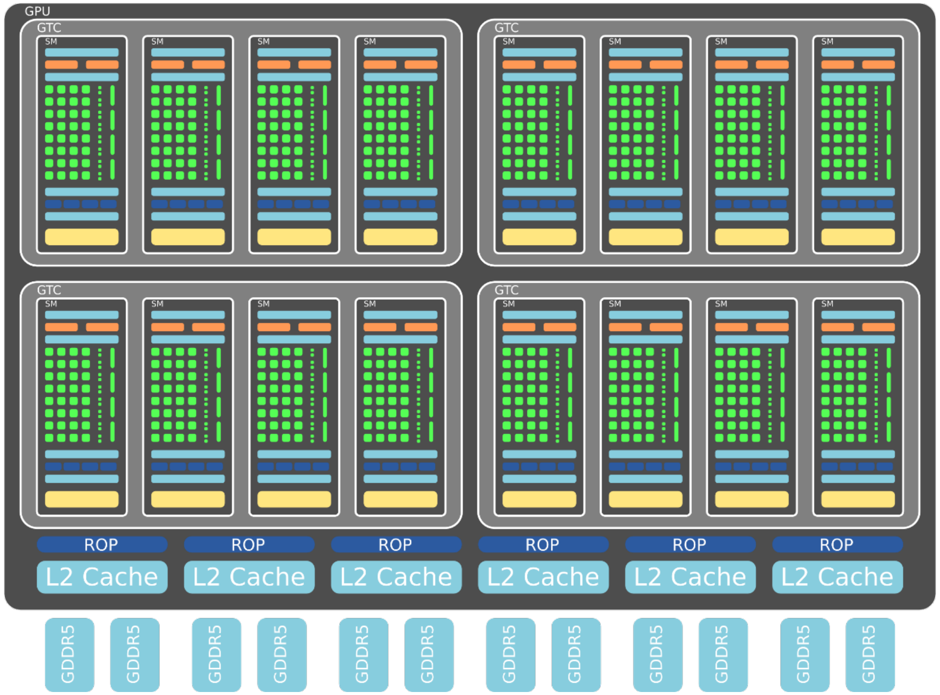
Computer Cluster

- Multiple stand-alone computers are connected by interconnection networks
- Each stand-alone computer is a multicore system with its own local memory
- The majority of supercomputers are a kind of computer clusters
  - Each compute node contains not only multiple cores, but also GPUs
  - e.g., The currently fastest supercomputer in the world “Frontier” consists of 9472 compute nodes, each being of one CPU of 64 cores and four GPUs



Modern GPU

- Each GPU contains
  - Multiple SMs – Streaming Multiprocessor
    - Multiple SPs – Streaming Processor (“scalar processor core”) (AKA “thread processor”)
    - Register file
    - Shared memory
  - Constant cache (read only for SM)
  - Texture cache (read only for SM)
  - Device memory



Modern GPU

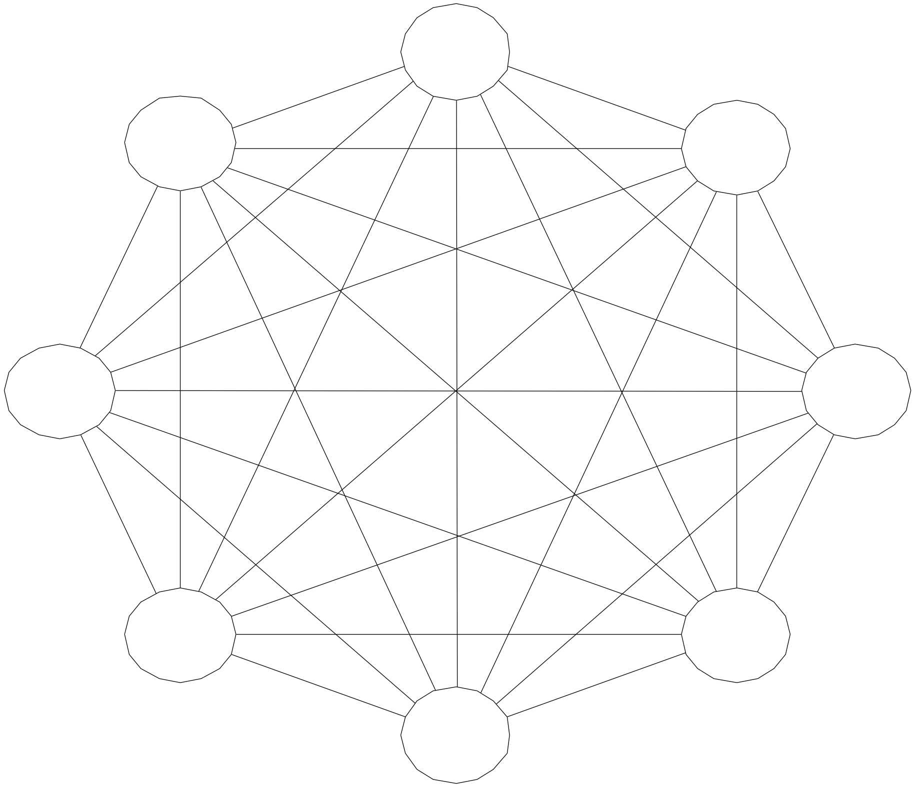
- new generation GPU, e.g., GeForce RTX 4090:
  - Each streaming multiprocessor (SM)
    - 128 cores (streaming processors, or SPs)
    - 16MB of shared memory
    - 32MB register file
    - 4 tensor cores for machine learning
  - The chip has 128 SMs, or 16,384 cores, 512 tensor cores and also 72MB L2 cache

Interconnection Networks

- The interconnection network is one of the most critical components in computer systems and it can have significant impact on the performance of applications, especially in large scale computing systems
- Two type of networks: static and dynamic
- In a static network, messages must be routed along established links
  - This means a single message must hop through intermediate processors on its way to its destination
- Dynamic networks establish a connection between two or more nodes on the fly as messages are routed along the links and switches

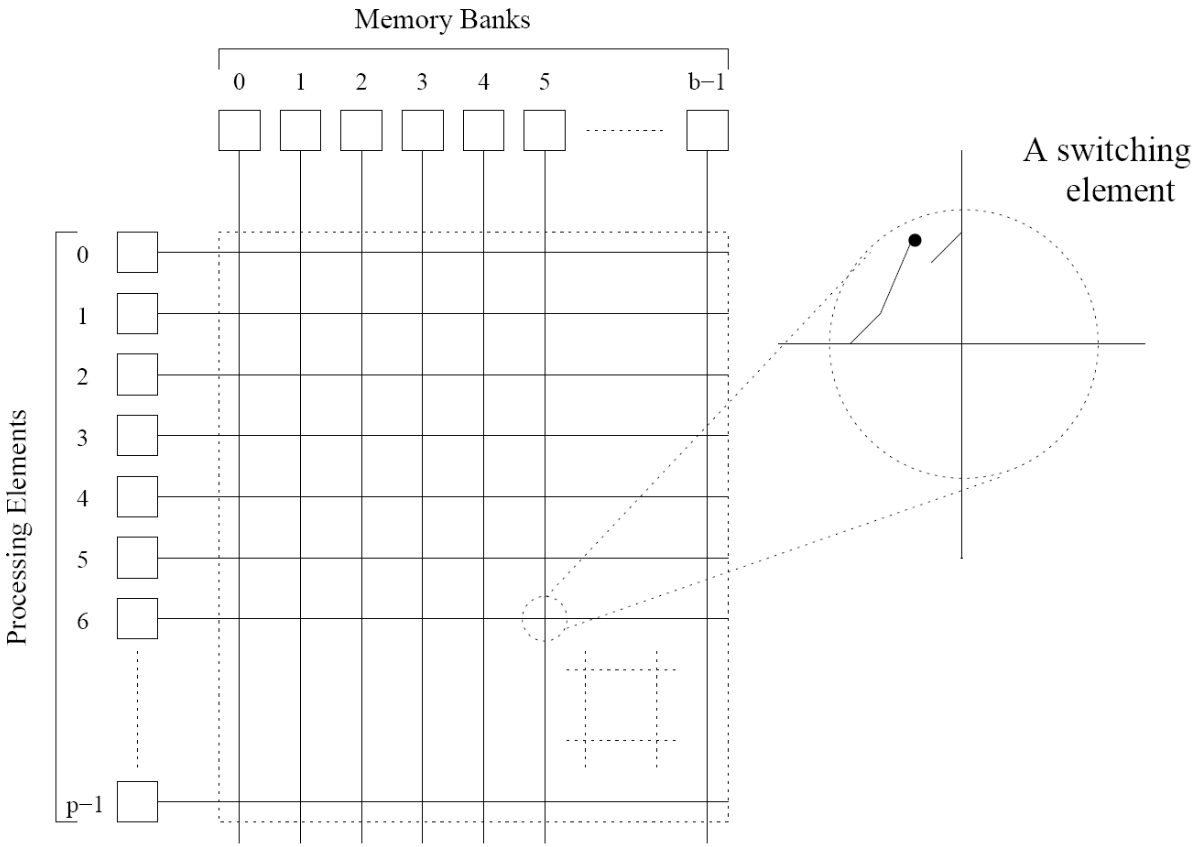
Interconnection Networks

- Completely connected network is a static network
- Each processor is connected to every other processor
- The number of links in the network scales as $O(p^2)$ where p is the number of processors
- While the performance scales very well, the hardware complexity is not realizable for large values of p



Interconnection Networks

- Crossbar is a switched network
- The cost of a crossbar of p processors grows as $O(p^2)$
- This is generally difficult to scale for large values of p



Interconnection Networks

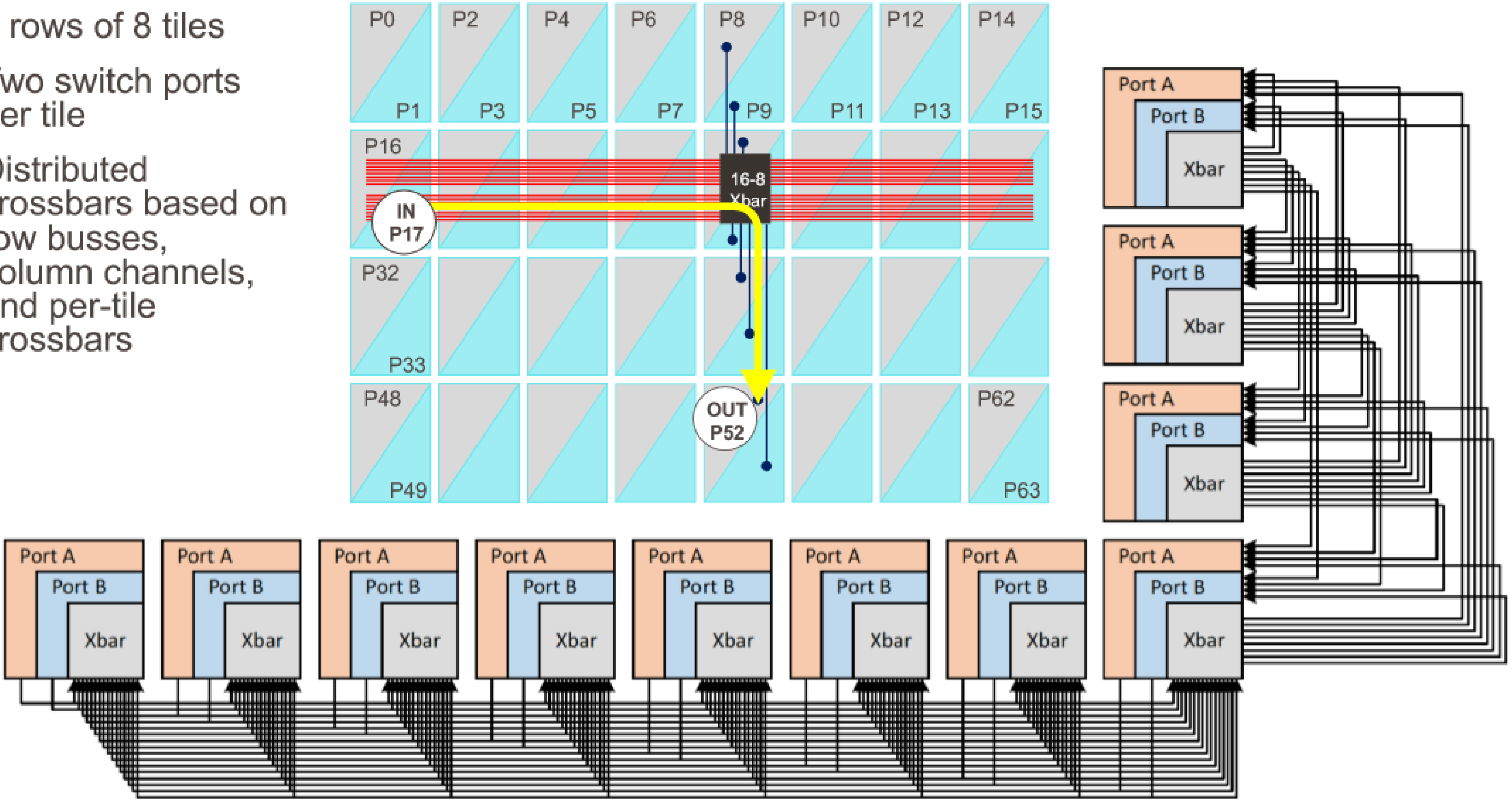
- Frontier is the world's first exascale supercomputer (10^{18} FLOPS), hosted at the Oak Ridge Leadership Computing Facility in USA
- It uses 9,472 compute nodes
- Each node consists of one 64 core CPU and 4 GPUs (thus a total of 606,208 cores and 37,888 GPUs)
- Computing nodes are interconnected using the state-of-the-art Slingshot network
  - The core of Slingshot interconnection network is the Rosetta switch
  - The default topology is Dragonfly – a hierarchical direct topology

Interconnection Networks

- Rosetta switch has 64 ports at 200 Gb/s which connect either compute nodes, or other switches to form different interconnection networks
- Ports are grouped into 32 tiles, with 2 ports per tile
- Tiles are organized in 4 rows of 8 tiles
- The tiles on the same row are connected through 16 per-row buses
  - Row bus used to send data from the corresponding port to the other ports on the row
- The tiles on the same column are connected per-tile crossbars
  - The per-tile crossbar has 16 inputs from the 16 ports on the row and 8 outputs to the 8 ports on the column
- It takes a maximum of 2 hops from one port to another (diameter = 2)

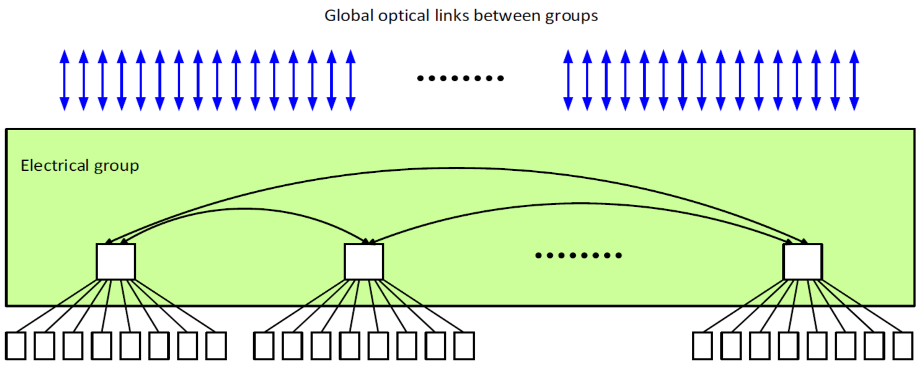
Interconnection Networks

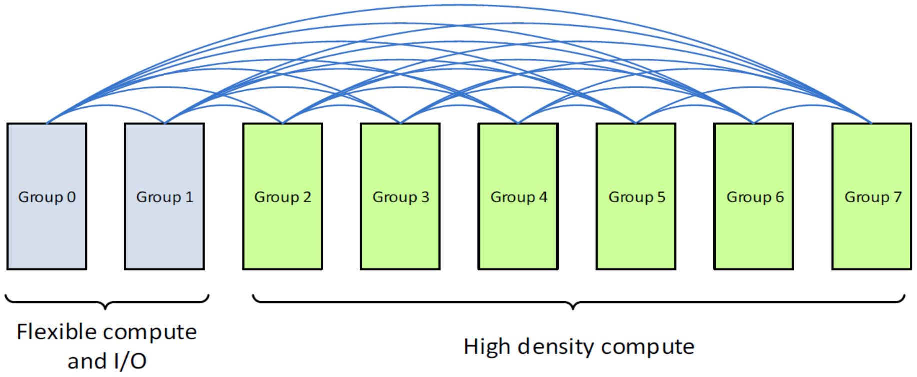
- 4 rows of 8 tiles
- Two switch ports per tile
- Distributed crossbars based on row busses, column channels, and per-tile crossbars



Interconnection Networks

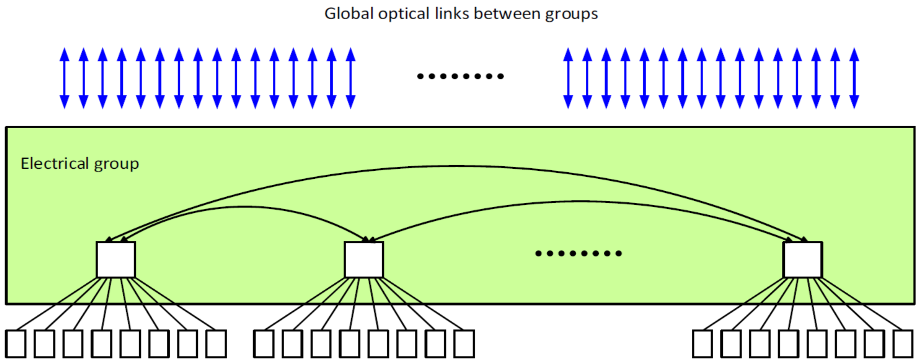
- Dragonfly is a hierarchical direct topology
  - Switches are organized into groups
    - Usually cabinets
  - In each group switches are connected in a fully (or completed) connected graph using electrical links (copper cables)
  - Groups are also connected in a fully connected graph, but using optical links (optical cables)

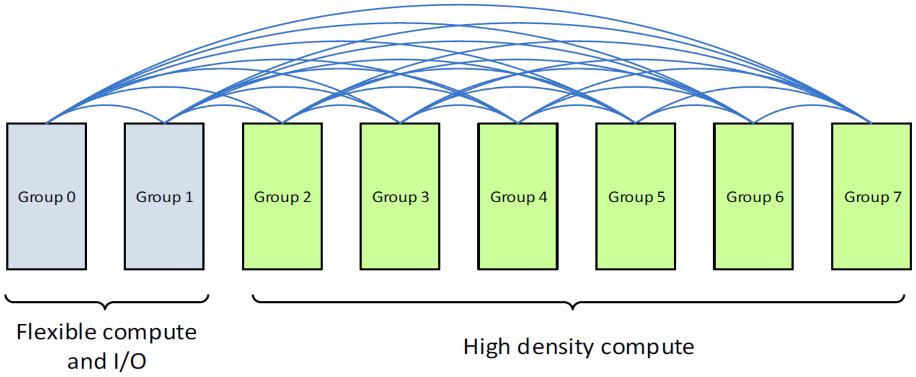




Interconnection Networks

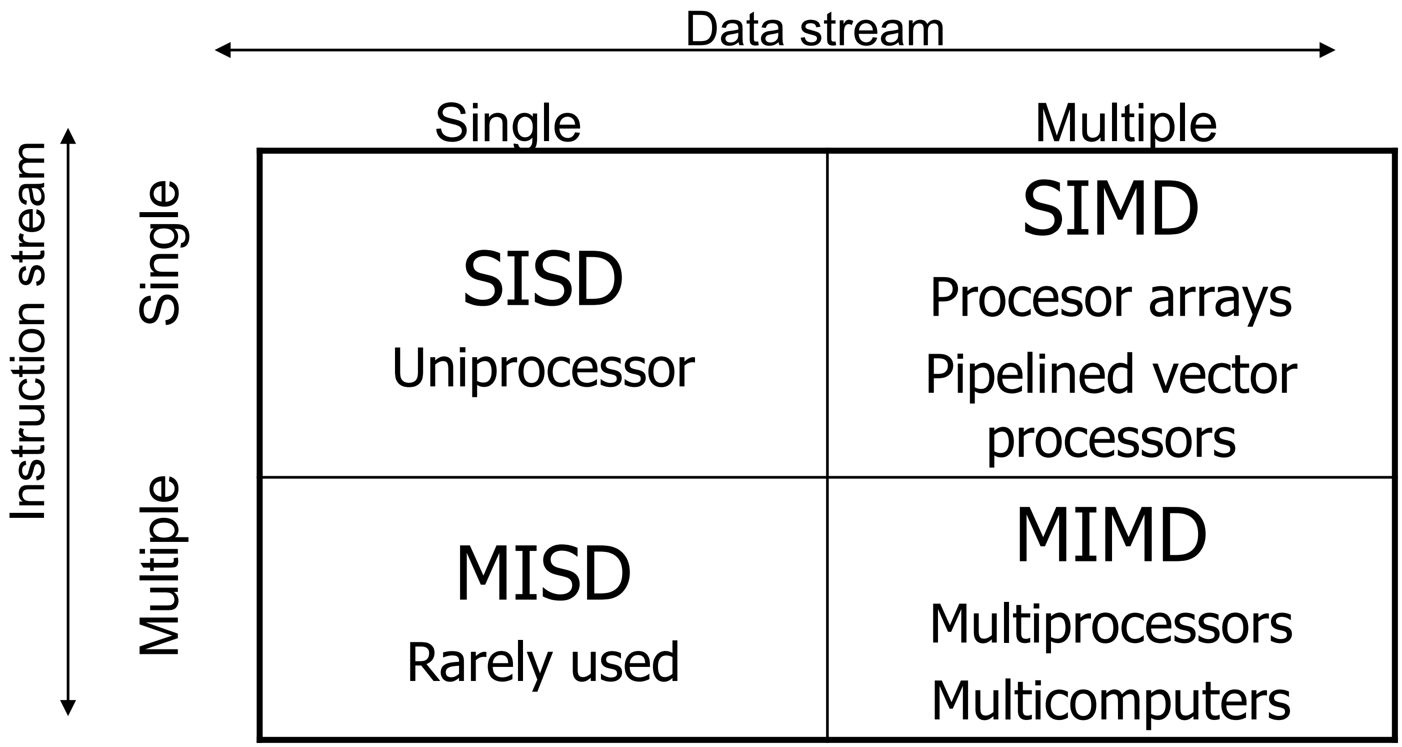
- Advantages:
  - Low latency: Due to the full-connectivity both within the group and between groups, Dragonfly has a diameter of 3 switch-to-switch hops
  - Low cost: Minimized long optical cables compared with other topologies for large systems
  - Highly modular and scalable





Computer Classification

- Flynn's Taxonomy: Prof Michael Flynn (Stanford University) proposed the method to classify computers in 60's
- Most modern computers are combinations of these



Data stream

Instruction stream		Single	Multiple
	Single	SISD Uniprocessor	SIMD Procesor arrays Pipelined vector processors
	Multiple	MISD Rarely used	MIMD Multiprocessors Multicomputers

Computer Classification

- Physical organization
  - Currently, parallel computers can basically be classified into three categories
    - Shared-memory machines (MIMD)
      - E.g., multicores with shared memory
      - Small scale, high end single node computing servers contain only several dozens of cores
    - Distributed-memory machines (MIMD)
      - Multiple computing nodes, each having its own local memory
      - Highly scalable
      - All supercomputers are distributed-memory machines
    - Accelerators
      - SIMD (single instruction multiple data) for data parallel computation
      - E.g., GPU, TPU
  - Large scale modern parallel computers are a combination of the above, e.g., a computer cluster consists of a number of computing nodes with distributed memory, and each node contains multiple cores with shared memory and also possibly one or even several accelerators

Parallel Programming Models

- Logical organization
  - (or parallel computing platform) – provides a way to think about organization of parallel programs
- Based on the classification we have three parallel computing platforms:
  - Shared memory
  - Distributed memory
  - SIMD (data parallel) and multithreading
- In this summer school
  - we shall discuss how to program
    - Shared-memory platform with OpenMP
    - Distributed-memory platform with MPI
  - we'll also briefly discuss CUDA programming for GPUs, but no exercises

Lab exercise 1: EduCoder (头歌平台)

- Get familiar with EduCoder
- Compile and run two “Hello World!” programs
  - omp_hw.c
    - Compile: gcc –fopenmp –o omp_hw omp_hw.c
    - Run: ./omp_hw
  - mpi_hw.c
    - Compile: mpicc –o mpi_hw mpi_hw.c
    - Run: mpirun –np x mpi_hw
      - x is an integer number, denoting how many processes will be created
      - You may run program several times with different x

Q&A

THE UNIVERSITY OF SYDNEY