BITS Pilani

WILP

AIML CLZG516

ML System Optimization

Shan Sundar Balasubramaniam

innovate achieve lead

BITS Pilani

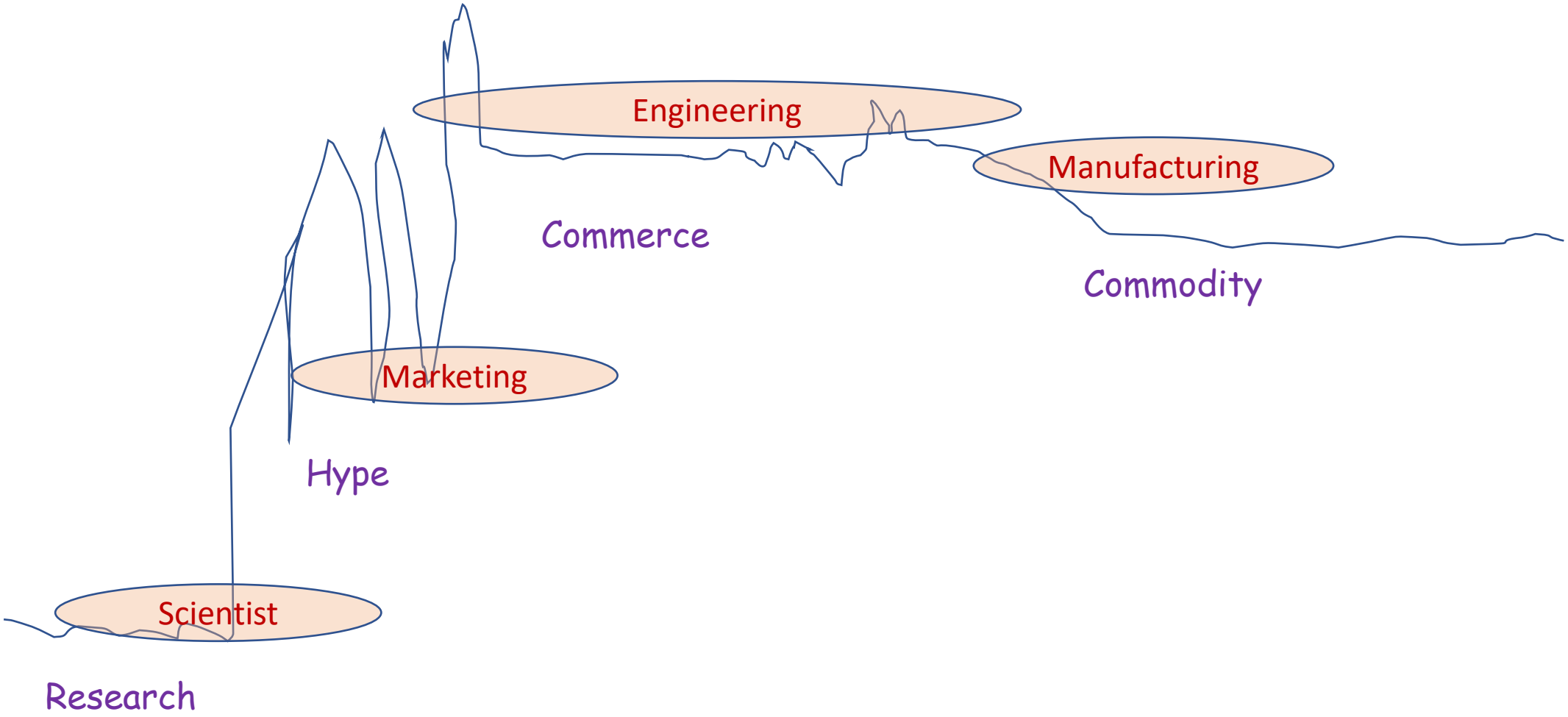
WILP

AIML CLZG516
ML System Optimization

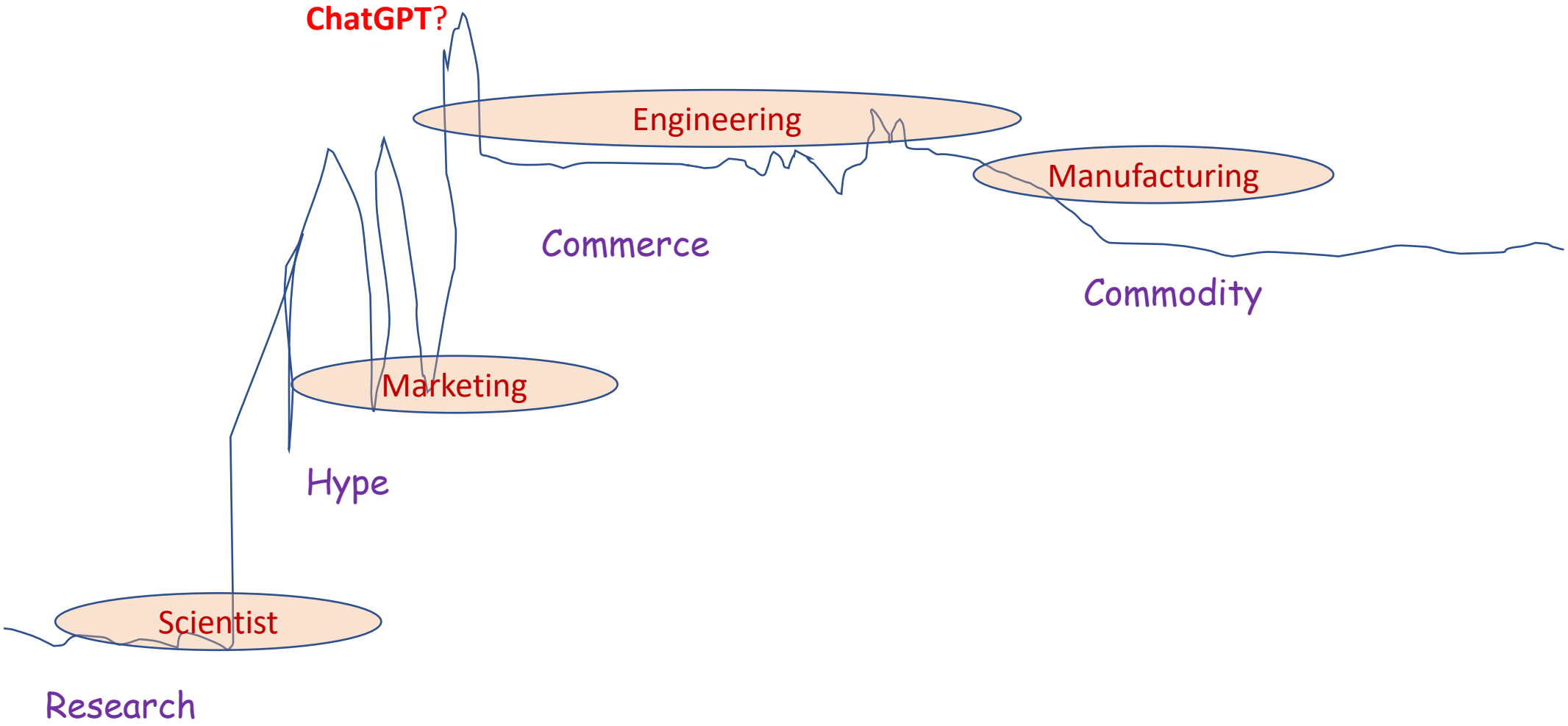
Session 1: 21 May 2023

Orientation: Course Introduction

Lifecycle of New Technologies



Lifecycle of AI/ML - Where are we ?



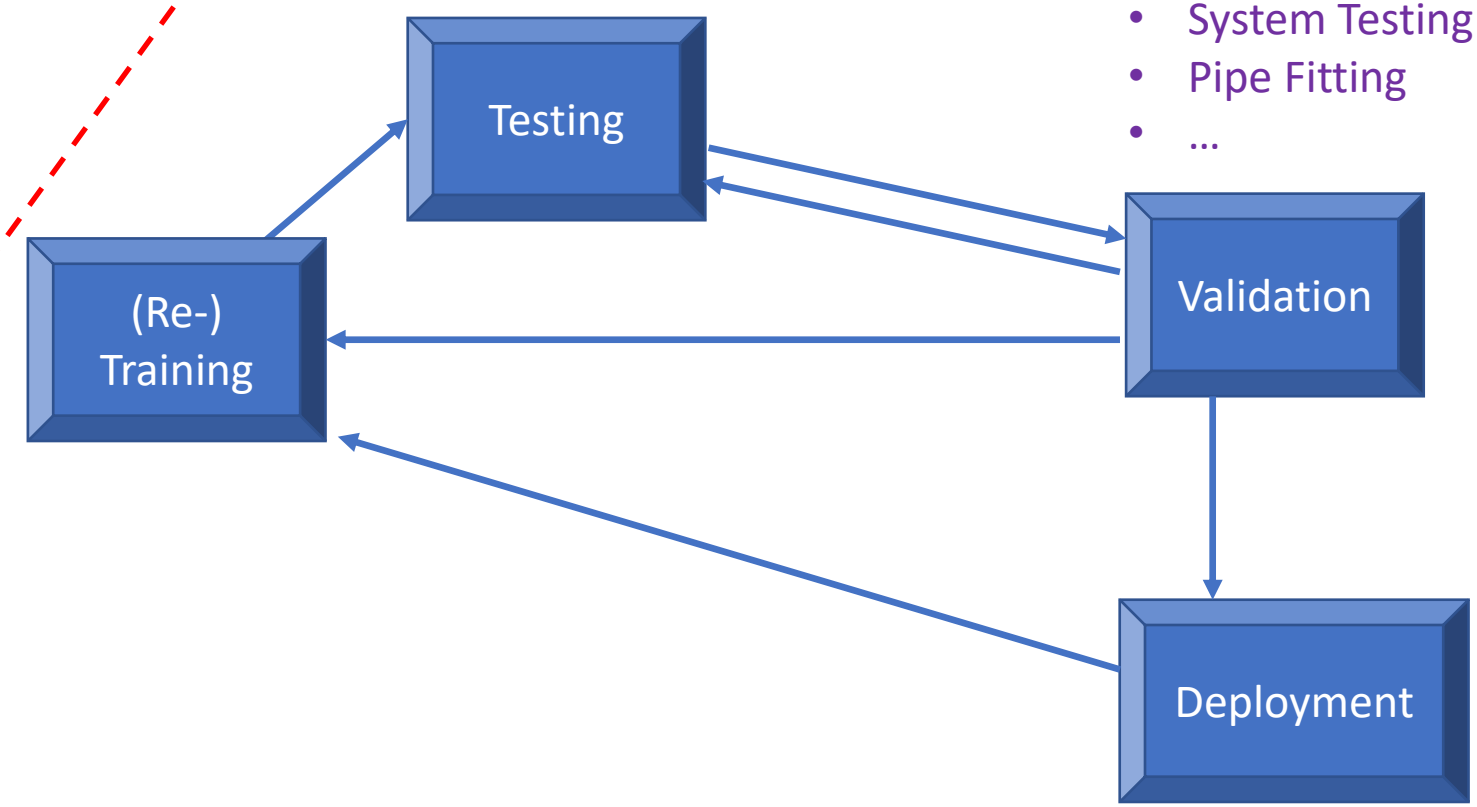
Machine Learning – Enterprise Practice

- AI and Machine Learning is becoming central to organizations:
  - No longer a one-off activity
  - Multiple problems / perspectives addressed through ML
  - Multiple ML solutions deployed
- ML is becoming a continual activity:
  - Data change; Context changes
    - Drift in the solution
  - Problems change; Requirements change;
    - New model(s) required
  - World changes; Expectations change
    - Performance and Standardization critical ==>
      - Packaging vs. Pricing

Operationalizing AI/ML

- Class-room View :
  - Development
    - Model Building / Training
  - Deployment
    - Inference

- Enterprise View:



- Compliance
- System Testing
- Pipe Fitting
- …

Extreme Scenario:

Data is arriving piecemeal (or streaming) ==> Training is incremental!

Operationalizing AI / ML: Cost

- Cost:
  - Time and Resources during Training vs Inference
- During Training:
  - Running Time of an algorithm:
    - E.g. k-means is an O(N*N) algorithm given N data points
    - E.g. SVM has a time complexity between $O(d*N^2)$ and $(d*N^3)$ where
      - d is the number of dimensions (of the data points) and
      - N is the number of data points

Cost during Training

- Example
  - E.g. SVM has a time complexity between $O(d*N^2)$ and $(d*N^3)$ where
    - d is the number of dimensions (of the data points) and
    - N is the number of data points
  - For a large dataset N, say, $N = 10^9$ and d=5 this could be costly:
    - Assuming 2 simple arithmetic operations per data point:
    - this amounts to at least 10^{19} operations
  - Given a 2.5 GHz processor, i.e. 0.4ns clock cycle
  - and 1 CPI (i.e. cycles per instruction), a measure of processor throughput
    - [simplistic but close to reality!]
  - 10^{19} operations will take close to 5.3 years
- Reducing running time during training is a big focus in this course!

Reducing running time

- Typical methods:
  - Parallelize or distribute computation:
    - Multi-threaded programming on multi-core processors
    - Massively multi-threaded programming on many-core GPGPUs
    - Distributed Programming on Scale-out Clusters of CPUs or GPUs
  - Hand-tuning or compiler-performed code optimization
  - Process = Program + Address Space (at run time)
    - Threads share address space:
      - Each thread gets its own call stack
      - Heap and global area are shared by all threads
      - Threads run on a shared memory model (e.g., multi-core, many-core processors)
    - Distributed programming is on Distributed memory ie. Memory of multiple computers (Processor+memory+disk+OS)

Cost during training

- Megatron-Turing NLG:
  - 530 billion parameters
- Microsoft and Nvidia claim to have used hundreds of DGX A100 servers
  - Each server costs ~200,000 $
    - Add the networking cost, the infrastructure cost is ~100M$
  - Each server consumes 6.5kW of power
    - Add a comparable cooling cost!
- We will NOT do much about power consumption in this course!
  - But we will look at reducing model size as an important aspect!

Model Size

- LLMs (Large Language Models) like GPT-3 and Bard are notoriously large.
  - But there are systematic approaches to reduce model size
    - Without compromising the accuracy too much.
  - We will look at model compression in this course!

Cost during Deployment

- When a model is deployed:
  - Requests come in and the model responds with inferences
    - E.g. if your model is a classifier:
      - For a new input x,
      - the response is C(x) such that $x \in C(x)$
- Performance Parameters for this phase:
  - Throughput:
    - Number of inferences over a unit of time
  - Response Time: Time take to serve one inference
    - Consider the classifier example with a (data) cluster example!
      - Will there be a difference in response time?

Deployment Range

- The model (that has been trained) or an application using the model could be deployed on a variety of platforms:
  - A server (or a workstation)
    - What if the model is large?
  - The Cloud
    - The cloud can provision large infrastructure to host a large model:
      - Increase the number of servers hosting and accepting requests thereby improving throughput and response time!
        - But there is always delay
          - i.e., network latency in reaching a remote server or a server on the cloud (and getting the response back)
  - A mobile phone:
    - best end-user response time but cannot host large models.

Sequential vs. Batch

- BATCH (Assumption): Requests are collected together and sent
- Responses are collected together and sent

- Ans.
- Part (A) If the model server is parallel multiple threads or processes could respond in parallel thereby improving response time and thru'put.

- Part (B) [Always] Messaging/Communication cost may be reduced:
- Communication cost = setup-cost + transmission cost
  - set-up cost is fixed per message
  - Transmission cost is proportional to the length of the message

Content & Pedagogy

- Focus on systems, programming techniques, and analysis
  - Pragmatics and Implementation to be learnt by doing – enabled by Assignments and Project.

- Lecture Sessions are expected to be interactive:
  - students are expected to raise questions and
  - the instructor will ask questions (which the students are expected to answer)

Evaluation

- Mid-term test and final exam – centrally scheduled by BITS
  - A Total of 55% weight = 25% for test + 30% for final exam
- 1 Assignment and 1 Project : Team exercises, Take-home
  - (15+30 =) 45% weight

Assignment and Project

- They are meant for learning
  - Expected:
    - One complex-end-to-end piece of optimization completed
    - One cutting-edge optimization technique learnt
  - Team-work with identifiable and quantifiable individual contributions
    - Evaluation both at team level and individual level

BITS Pilani

WILP

AIML CLZG516

ML System Optimization

Shan Sundar Balasubramaniam

innovate achieve lead

BITS Pilani

WILP

AIML CLZG516
ML System Optimization

Session 2: 28 May 2023

Parallel Programming Models:

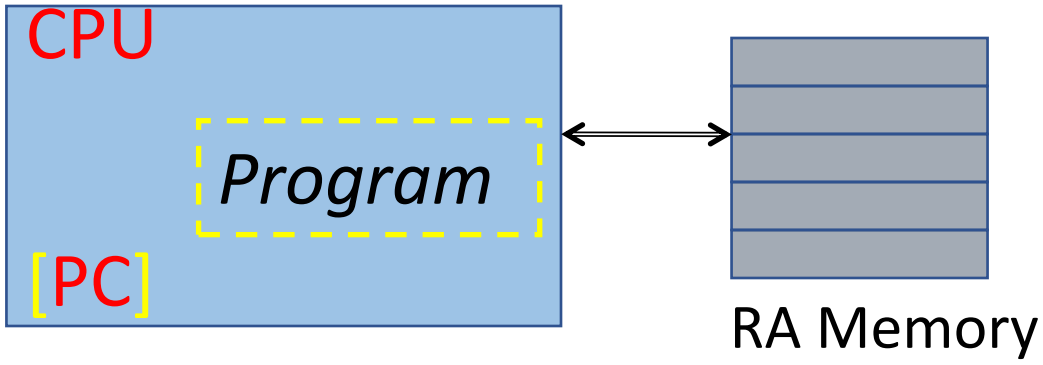
- Pipe-lined, Data-Parallel, Task-Parallel, and Request-Parallel
- Speedup

Algorithm Design - Sequential

- Generic Machine Model
  - Random Access Machine Model

Typical Instructions

- Arithmetic / logic operations,
- Load / Store, and
- Jump / Branch



PC: Program Counter
(tracks the next instruction to be executed)

RAM: Random Access Memory
(*cost of access is uniform across locations*)

Executing an Instruction

- Different stages of Instruction Execution:

Fetch Instruction	*Move the next instruction (tracked by PC) to a register*
Decode Instruction	*Identify Operator and (data) addresses*
Load (data)	*Move data into register (if needed)*
Execute	*Perform the operation*
Store (result)	*Move the resulting data to memory*

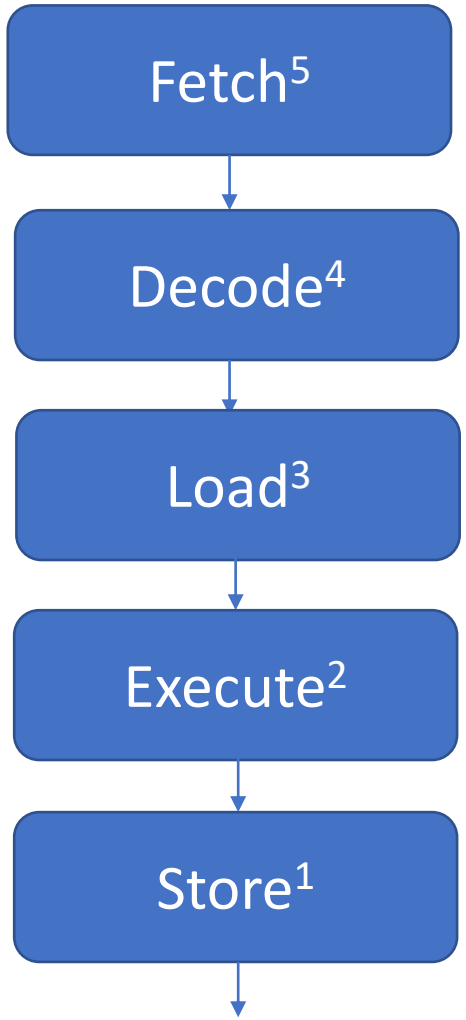
If separate circuitry is designed for each stage-
so that the stage take the same amount of time -
then a sequence of instructions can be executed in a pipeline

Instruction Pipeline

Given a sequence of instructions of the form:

- I1
- I2
- I3
- I4
- I5
- …

execution in a pipeline would appear thus ==>



If each stage takes 1 clock cycle, then throughput has increased :

- from 1 instruction per 5 cycles (sequential)
- to 1 instruction per 1 cycle (pipelined)

Q: What about Turn-around-Time aka response time?

Modern Processors

- Modern processors (since Intel Pentium circa 1991)
  - typically include a pipeline that is several stages (>5) deep
- Throughput in processors is measured in CPI (or Cycles Per Instruction):
  - For an ideal pipeline design: CPI is 1
  - In practice, it may be more (Why?)
    - but on an average it is kept close to 1

Pipeline Throughput

- Speedup (i.e. throughput increase) is k for a k-stage pipeline
- Factors that may slow down the pipeline:
  - Some stage(s) take more time than others
    - Q: What is the impact on CPI if one stage takes 10% extra time?
  - Memory access takes more time (i.e., LOAD and STORE)
- Modern processor pipelines are designed
  - such that all stages take almost the same time (except for LOAD and STORE)

Pipeline Throughput and Memory Access [2]

- Memory access is slower compared to Processor speed:
  - Typical processor clock cycles
    - e.g. 2 to 3 GHz
    - i.e., in-processor operation may take only 0.33 to 0.5ns
  - Access from Memory (DRAM) will take around
    - 50 to 200ns
  - Access from Cache (SRAM) – if available – may take
    - 5 to 10ns.
- Modern architectures use multiple levels of caching and other techniques to keep the access time low.
- Compiler and processor collaborate to keep the frequency of memory access operations low.

Pipeline Throughput and Memory Access

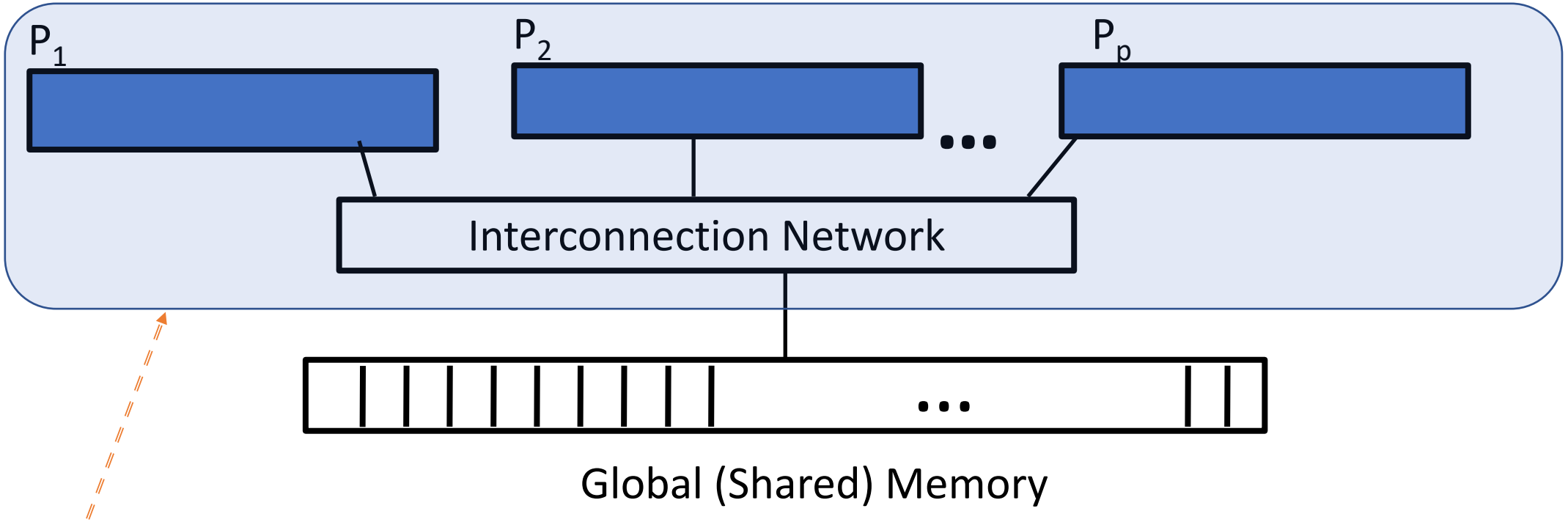
- STORE operations may be executed *asynchronously*:
  - i.e. processor does not wait for data to be stored in memory
    - Store buffers (i.e., buffer registers inside the processor) are used to store the data temporarily
    - while data is transferred to memory without processor involvement

Software Pipelining

- The idea of a pipeline can be extended to Software Design:
  - Break a long task into multiple stages
    - so that the stages take (roughly) the same amount of time.
  - If there is a stream of data to be processed by the data,
    - then the stream can be fed to the pipeline for improved throughput.
- We will revisit this later!

Algorithm Design – Parallel: *Shared Memory* Model

Target environment:



e.g. a multi-core chip

Typical Instructions

- Arithmetic / logic operations,
- Load / Store, and
- Jump / Branch

Multi-threaded Programming:

each thread runs on a separate core

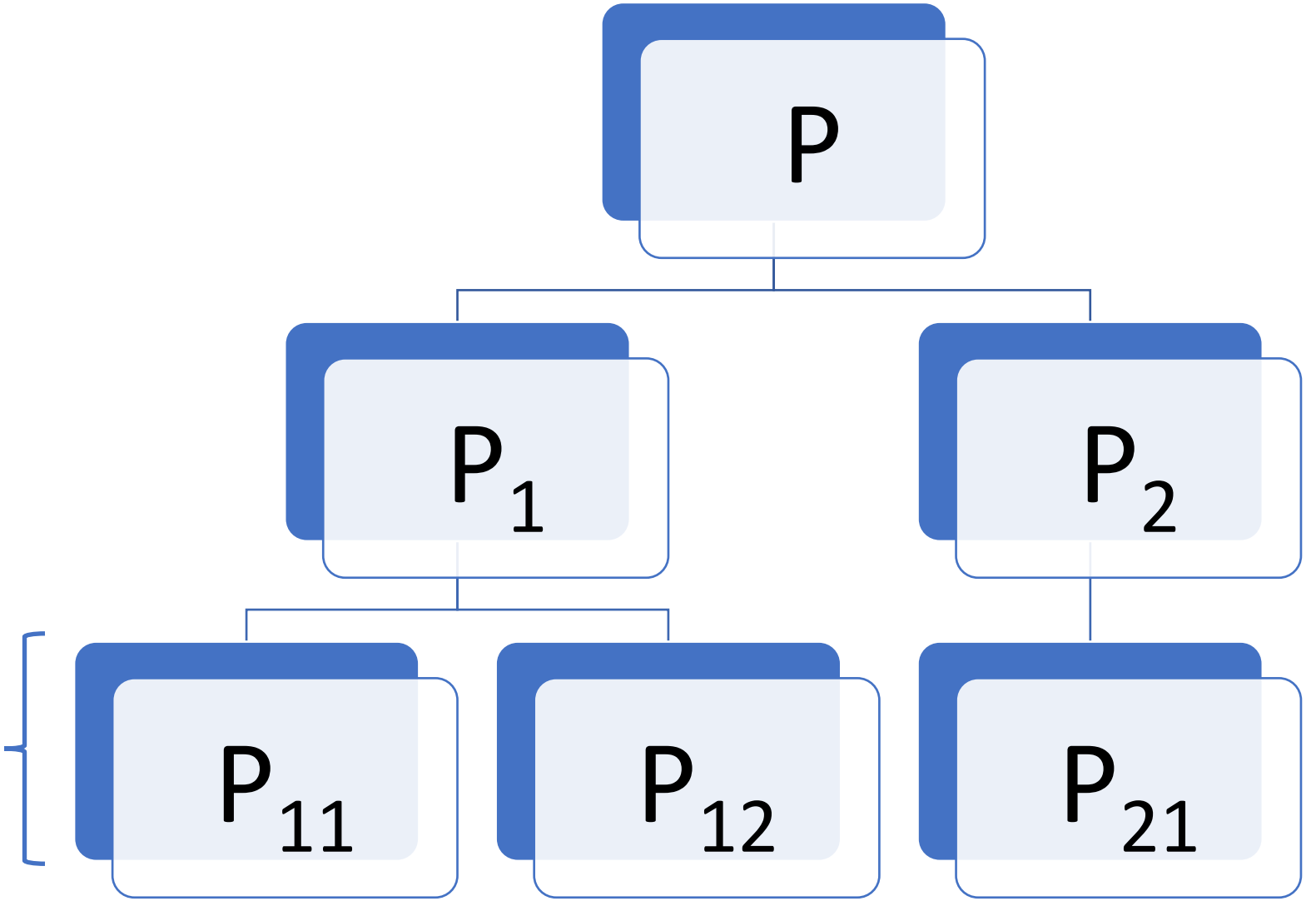
Algorithm Design

- Top-Down Design (Top Down Decomposition)
  1. Divide the problem into sub-problems.
  2. Find solutions for sub-problems
  3. Combine the sub-solutions.
- How do we find solutions for sub problems?
  - Apply top-down design recursively (i.e. divide each sub-problem further)
    - Q: When do we stop dividing?
    - A: When we reach "atomic" problems.
      - Atomic problems have known solutions
- Does any decomposition work?
  - Divide (the problem) only if you know how to combine (the solutions)

Top Down Design - Parallel

- Does any decomposition work?
  - Divide (the problem) only if you know how to combine (the solutions)
  - But also:
    - *Mapping sub-problems to processors*
    - *Where is the combination done?*
  - Number of sub-problems?
    - Processor utilization is the key!
      - i.e. More the number of processors more the number of sub-problems!

Top-Down Design - Parallel



P_1 and P_2 can run in parallel

P_{11}, P_{12} and P_{21} can run in parallel

Example: Search a key k in a list Ls of size N

Data: Assume **Ls[0..N-1] is stored in shared memory**

t is N/p

for processor P_j from j = 0 to p-1
    do res_j = search(k, Ls[j*t..(j+1)*t-1])

for processor P_0 :
do res = TRUE ;
for j = 0 to p-1 do res = res AND res_j

This is an example of data parallel programming!

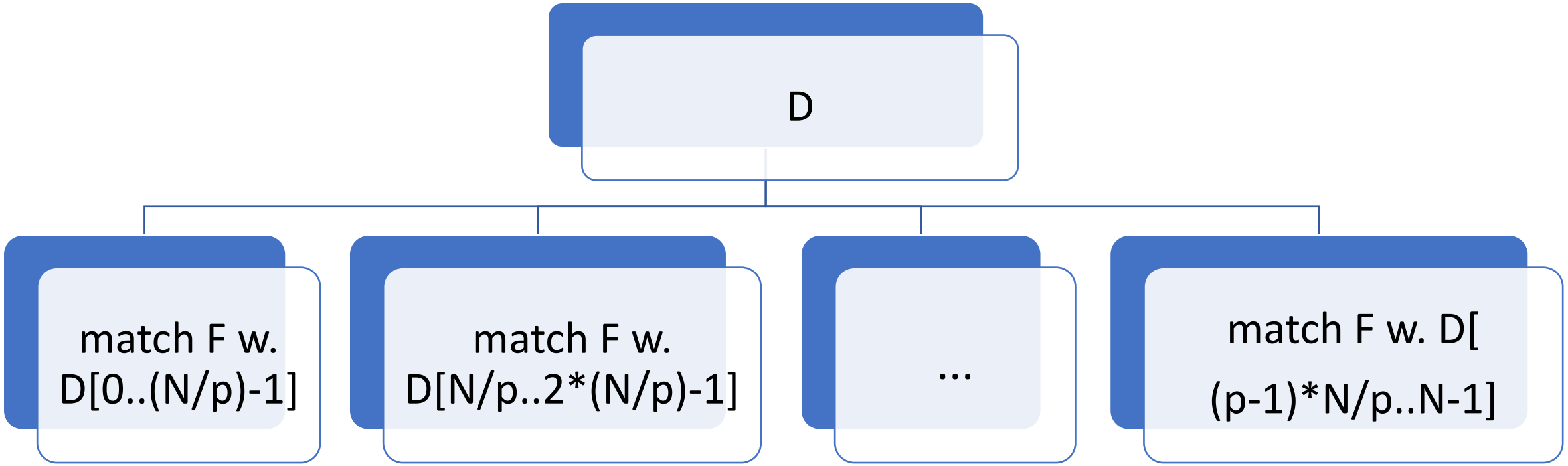
Data Parallel

- Data Parallel execution (or computation):
  - The same task executes independently (i.e., in parallel) on different data
    - i.e., divide given data into (roughly) equal-sized subsets
      - and the same task is replicated and run on different processors – one for each subset.
- Note that we are assuming a shared memory model
  - i.e., all processors can access the (global) shared memory
    - Dividing data may simply become setting (boundary) markers!
- This may result in memory contention:
  - i.e., performance may not scale (with number of processors)

Data Parallel Execution - Example

Fingerprint Matching:

- *Match a given print F with a database D of prints available*



Data Parallel Execution - Exercise

- Vector Product A.B for two vectors – each of length N
- $\Sigma_{j=1 \text{ to } N} A[j]*B[j]$
- N processors:
  - For each processor j=1 to N do: A[j]*B[j]
- How to do the addition? Can it be done in data-parallel fashion?
- p processors: Change the code!

SPMD

- Data-Parallel execution is also referred to as
  - Single Program Multiple Data (SPMD) programming (because a single program i.e. the same program) is executed on all processors
- This model Data-Parallel or SPMD is preferred where feasible
  - because of ease of programming and efficiency.
- In the parallel programming world, efficiency is measured as ***speedup***:
  - i.e., the ratio of time taken by a parallel algorithm to time taken by a sequential algorithm

Speedup

- Speedup (in running time) of a given algorithm *A* running on **p** processors is defined as:
  - Speedup(p) = (Time taken by *A* on 1 processor) / (Time taken by *A* on p processors)
- All parts of a program may not run independently or in parallel :
  - Memory contention
  - Data (structure) contention
  - Mutually exclusive access (e.g. update operations or transactions) of shared data
  - Data dependency (result of a task must be input to another)

Speedup – Amdahl's Law

- Assume that a fraction **f** of a task is not parallelizable (e.g., due to constraints seen in the last slide)
- **Speedup(p) = 1/(f + (1-f)/p)**
  - i.e., the parallelizable fraction (1-**f**) of the program has been sped-up by a factor of **p**, the number of processors
  - But the other part takes the same (fraction of) time **f**

- By definition, f=0 in data-parallel execution or an SPMD program:
  - and speedup(p) = p
- When speedup(p) is proportional to p, we say that the algorithm is **scalable.**

BITS Pilani

WILP

AIML CLZG516

ML System Optimization

Shan Sundar Balasubramaniam

innovate achieve lead

BITS Pilani

WILP

AIML CLZG516
ML System Optimization

Session 3: 4 Jun. 2023

Parallel Programming Models

[continued.]

- Map-Reduce Pattern
- Task-Parallel and Request-Parallel

Parallelization of ML Algorithms - Examples

Parallel Design Pattern map

- **map** is a data-parallel programming construct
- **map f Ls** expresses "execute **f** on all elements of **Ls**" using **p** processors
  - Where **Ls** has been distributed among the **p** processors (i.e. their memories)
- **map** is implicitly data parallel

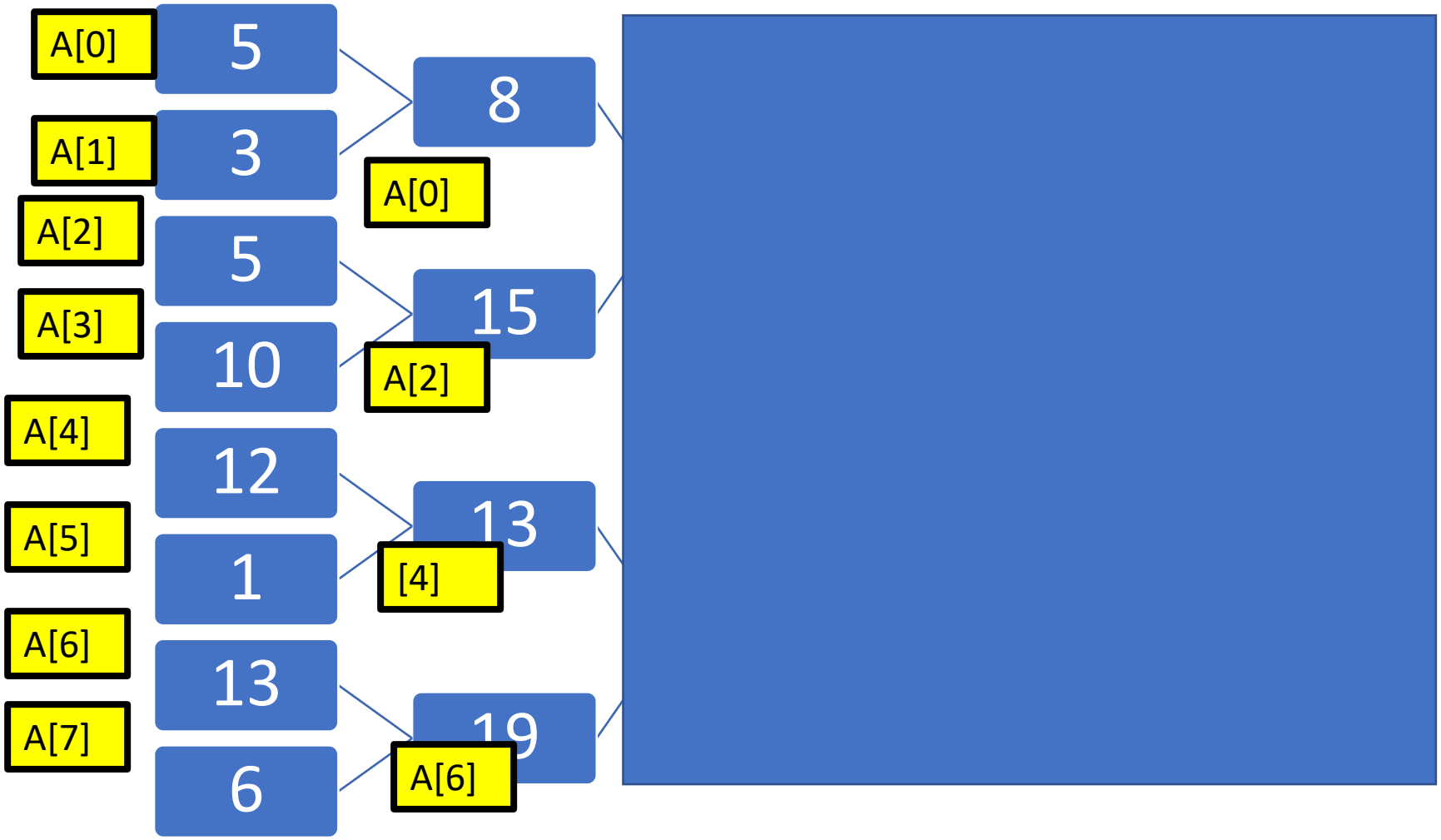
- Exercise:
  - Implement the data-parallel examples (from previous class)

Parallelization Constraints

- An algorithm may have inherently sequential steps that are not parallelized (or not fully parallelized) naturally.
  - Consider the problem of adding a list of numbers
    - This is an example of
      - (Inverse) Tree Parallelism or
      - The parallel design pattern ***reduce***

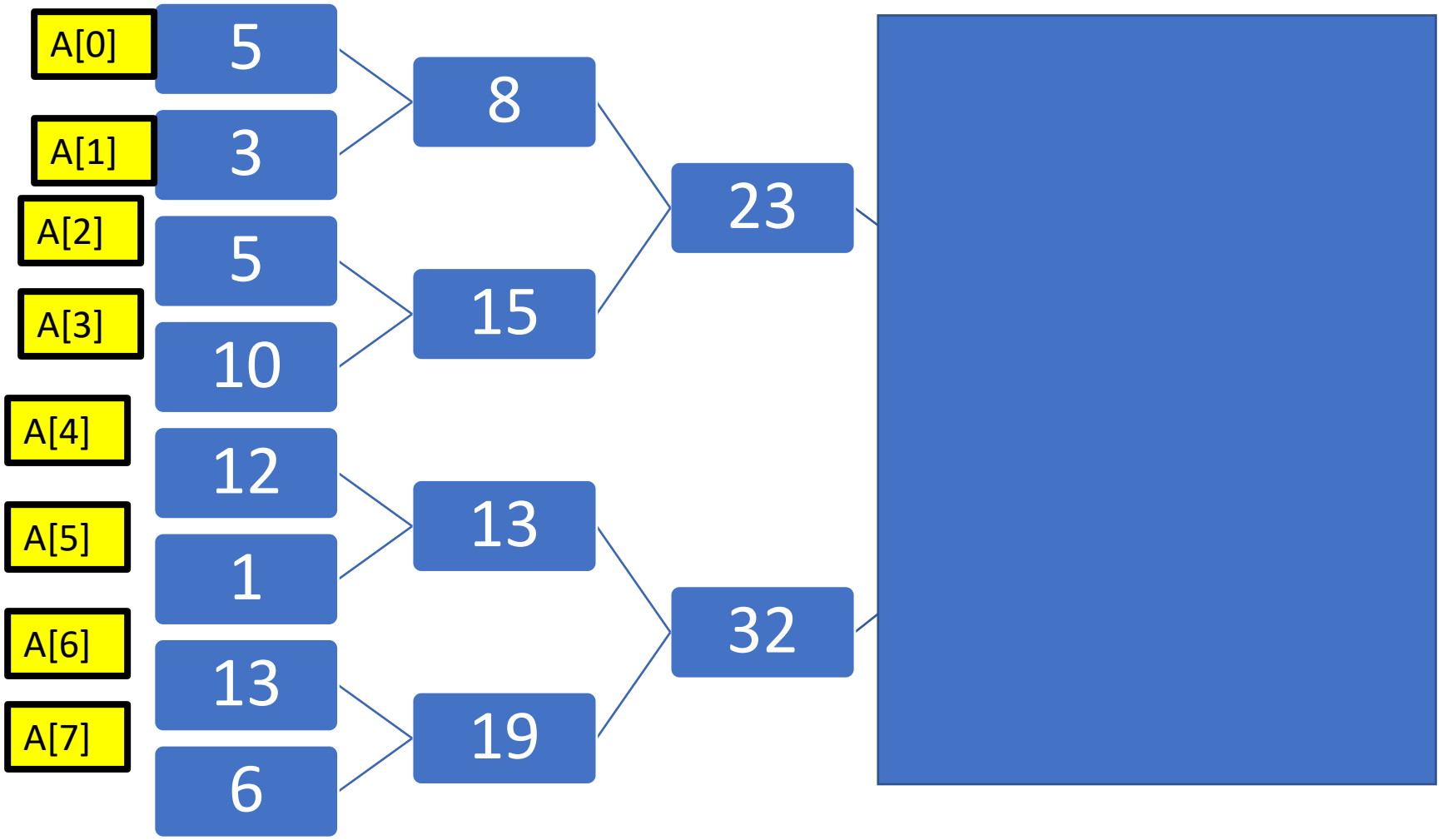
(Inverse) Tree-Parallel Algorithm : Summation

(Inverse) Tree-Parallel Algorithm : Summation



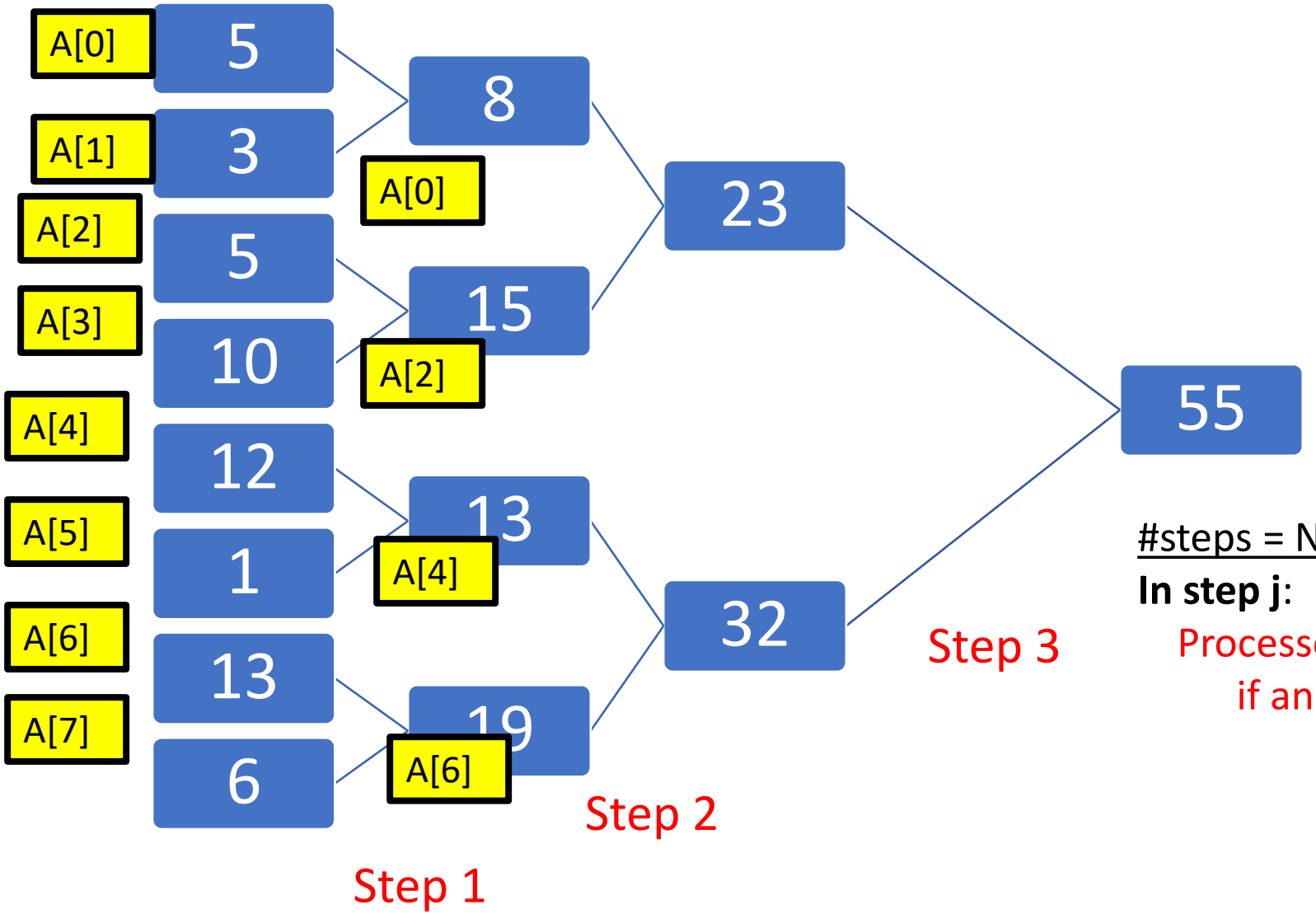
Step 1

(Inverse) Tree-Parallel Algorithm : Summation



Step 2

(Inverse) Tree-Parallel Algorithm : Summation

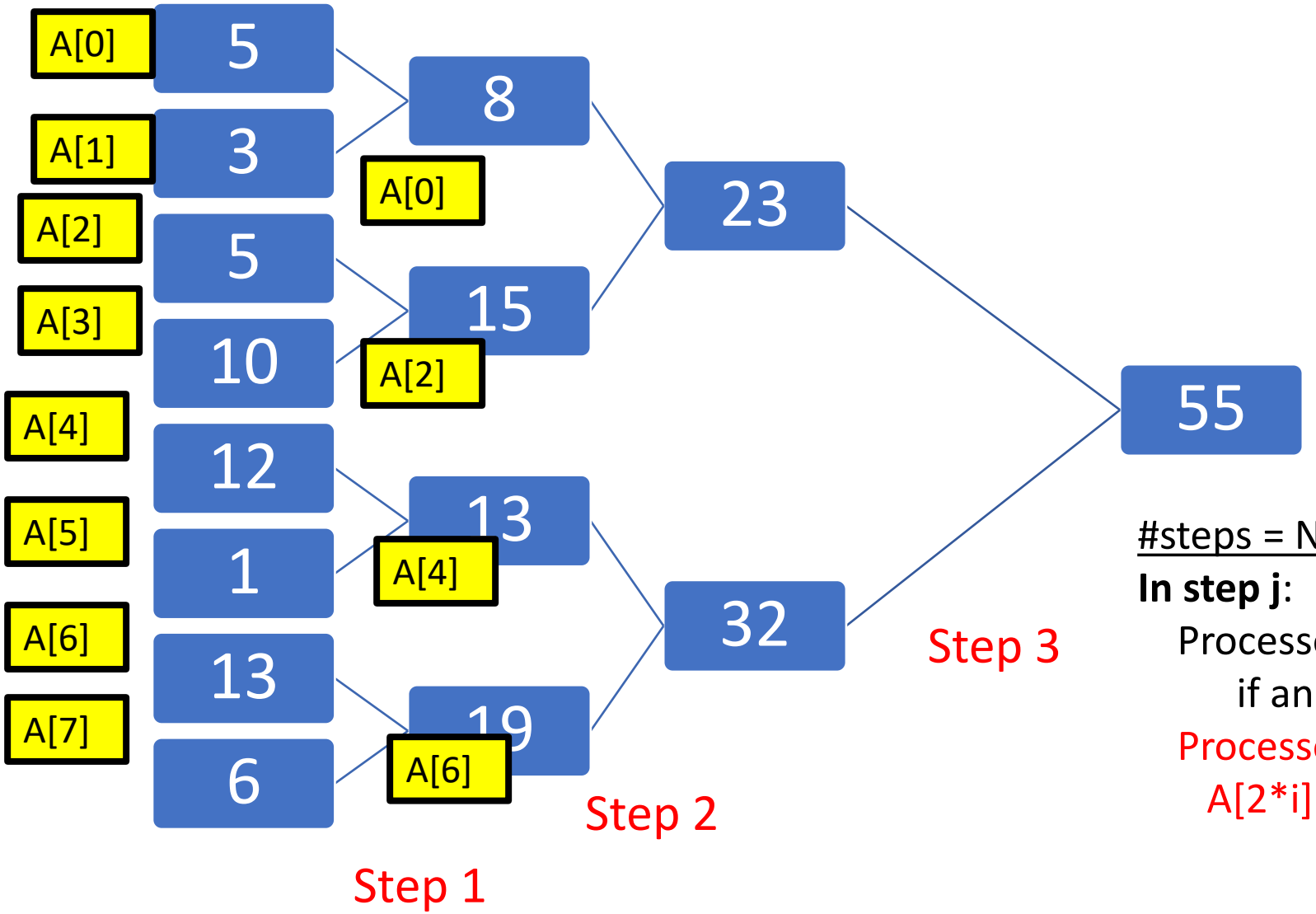


#steps = N for 2^N inputs

In step j:

Processor P_i participates if and only if $i\%2^j == 0$

(Inverse) Tree-Parallel Algorithm : Summation



#steps = N for 2^N inputs

In step j:

Processor P_i participates if and only if $i\%2^j == 0$

Processor P_i does:

$A[2*i] = A[2*i]+A[2*i+2^j]$

(Inverse) Tree-Parallel Algorithm : Summation

- Pre-condition: List A[0..n-1] in global memory
- Post-condition: sum A[0] in global memory
- Global variables: A, n, and j
- begin
  - for all Pi where 0 <= i <= floor(n/2)-1 {
  - for j = 0 to ceil(log(n))-1 {
  - if (i mod 2^j = 0) then
  - A[2*i] = A[2*i] + A[2*i + 2^j]
  - }}
- end

Decide which processors i participate in step j

Distance between self and (processor holding) the other data

(Inverse) Tree-Parallel Algorithm : Summation

- Precondition: List A[0..n-1] in global memory
- Postcondition: sum A[0] in global memory
- Global variables, A, n, and j
- begin
  - spawn (P0, P1, … Pk) where k=floor(n/2)-1
  - for all Pi where 0 <= i <= floor(n/2)-1 {
  - for j = 0 to ceil(log(n))-1 {
  - if (i mod 2^j = 0) and ($2*i + 2^j < n$) then
  - $A[2*i] = A[2*i] + A[2*i + 2^j]$
  - }}
- end

Boundary condition when n is not a power of 2

Algorithm : Summation - Performance

- Complexity of the algorithm:
  - Summation requires ceil(log n) steps for n inputs
    - Each step is $O(1)$ time
  - Total time $\theta(\log n)$ given n/2 processors
    - Compare with sequential algorithm
    - Speedup: $T_{seq} / T_{par} = T(n, 1) / T(n,p)$
      - Speedup(n/2) = (n-1) / log(n) = $O(n/\log n)$
        - This is less than ideal speedup!

Parallel Reduction - Template

Template REDUCE

- Precondition: Inputs, G, in global memory
- Postcondition: Result in G[0]
- Global variables: n and j, apart from G
- begin
  - for all Pi where 0 <= i <= floor(n/2)-1 {
  - for j = 0 to ceil(log(n))-1 {
  - if (i mod 2^j = 0) and ($2*i+2^j$ < n) then
  - G[2*i] = G[2*i] OP G[$2*i+2^j$]
  - }}
- end

Example Instances:

- Maximum
- Sum of matrices
- Intersection of sets

Reduce as a construct

- reduce '+ Ls
  - Returns a single value (the sum of all values in Ls)
- Reduce BOP Ls
  - Extends the binary operator BOP over a list of values
- This is valid only if BOP is associative
  - i.e. (x BOP y) z = x BOP (y BOP z)
- Examples
  1. Maximum of a list of values
     - BOP is *max*
  2. Sum of all matrices in a list
     - BOP is *matrix-sum*
  3. *Merge a list of sorted lists*
     - *BOP is (binary) merge*

Speedup(N/2) = N/log(N)

Exercise I: Vector Product

- Problem: Vector Product A.B for two vectors – each of length N
  - $\Sigma_{j=1 \text{ to } N}$ A[j]*B[j]
- Solution:
  - Step 1: N processors:
    - for each processor j=1 to N do: A[j]*B[j]
  - Step 2: N/2 processors:
    - Apply Reduction with * as the operator

This can be achieved using map-reduce:

- Make a list L of (A[j], B[j])
- L1 = map * L
- vp = reduce L1

Exercise II: Matrix Product

- Problem:
  - Multiply matrices $A_{m \times n}$ and $B_{n \times p}$
- Solution:
  - for each processor $P_{i,j}$ where (i = 1 to m) and (j = 1 to p)
    - C[i][j] = Compute vector product $A[i].B^T[j]$
      - where B^T is the transpose of B:
      - i.e., $B^T[j]$ is the j^{th} column of B

Express this using *map* and *reduce!*

Google's map-reduce framework

- Google's map-reduce has built-in capabilities for:
  - scheduling:
    - i.e., spawn processes depending on available processors
  - load-balancing:
    - i.e., move processes across processors to keep all processors equally utilized
  - fault-tolerance:
    - i.e., restart/resume processes that fail

map-reduce platforms

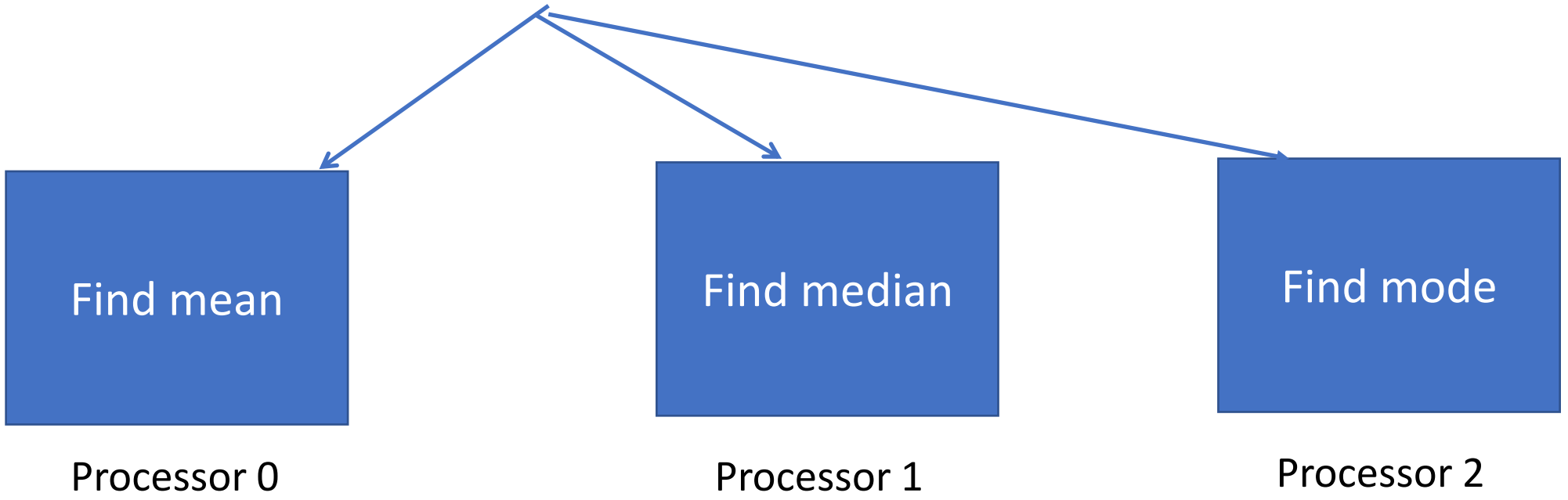
- Map-reduce is supported by different middleware platforms:
  - In particular, **Apache Spark** supports map-reduce on multi-core systems and clusters

- Exercise:
  - Install Apache Spark on your computer and
  - code the *matrix multiplication* example using map-reduce.

Task Parallelism

Task Parallelism – Example

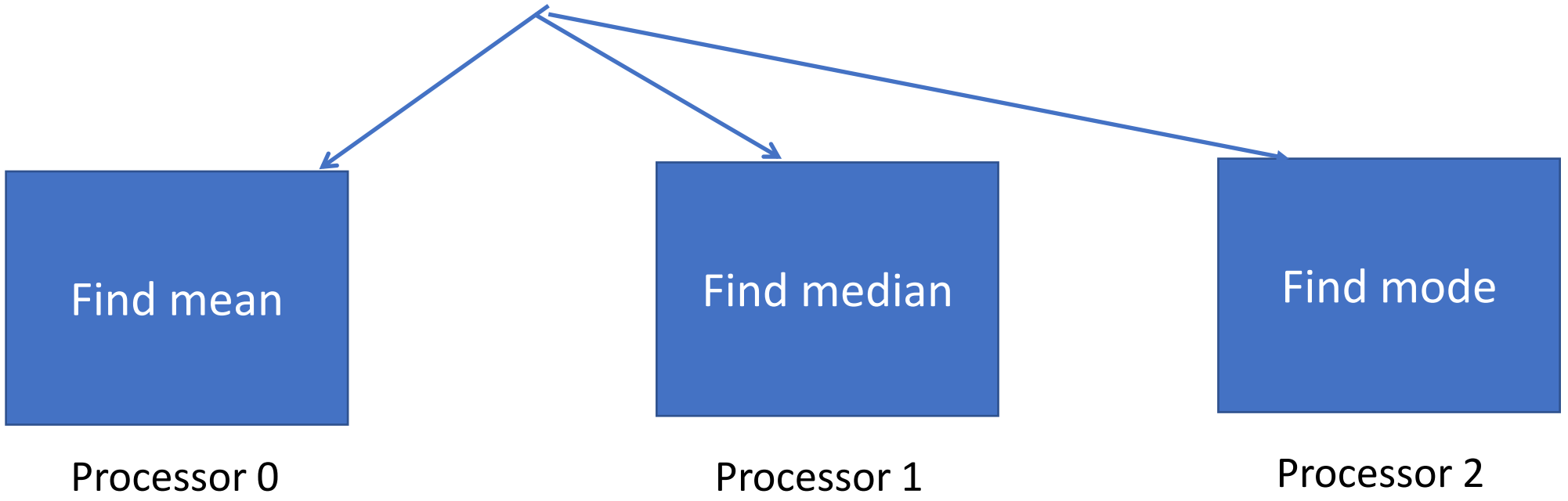
Problem: Given a list Ls of numeric values find the *mean*, *median*, and *mode*.



Task Parallelism – Example

Problem: Given a list Ls of numeric values find the mean, median, and mode.

Speedup = $T_{ser} / T_{par} = (T_{mean} + T_{med} + T_{mode}) / \max(T_{mean}, T_{med}, T_{mode})$

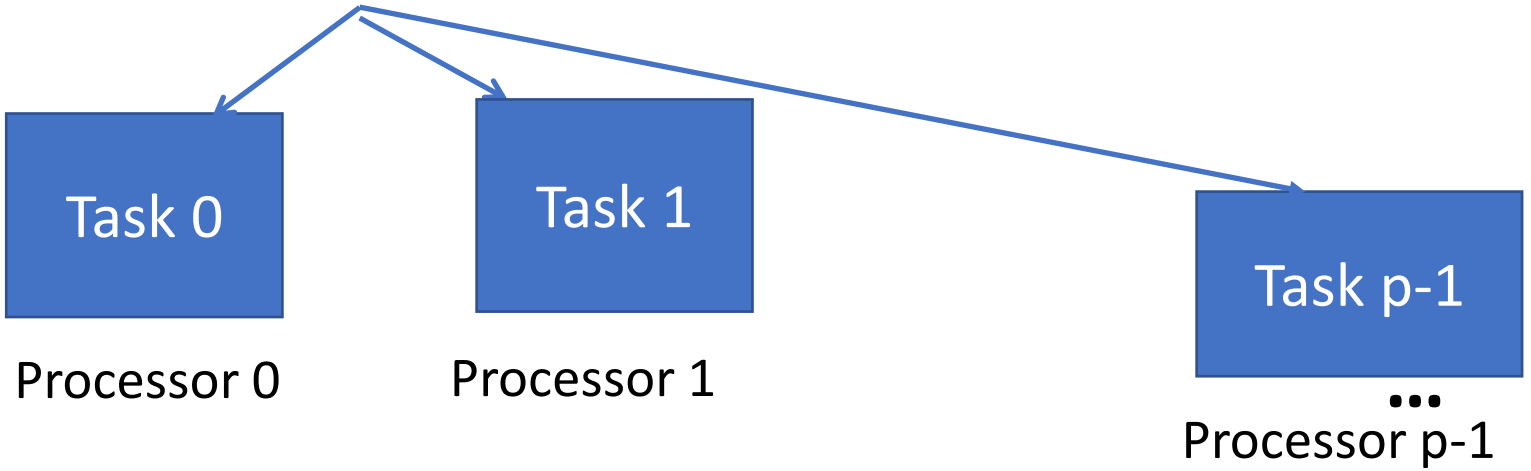


Task Parallelism

Speedup could be less than p because tasks could be uneven in size.

This is not scalable:

- if we put in more processors, we can't get more tasks running in parallel
  - because the number of tasks is fixed and/or small



Task Parallelism

- While task parallelism has speedup limitations, it is suitable for off-the-shelf computers:
  - Modern laptop or desktop computers and workstations are made out of a few (often one) multi-core chip(s):
    - The available parallel processing capacity is limited.

Programming Task Parallelism

- In shared memory computers:
  - Task parallelism can be implemented using multi-threaded programs:
    - One task per thread
      - Where data is stored in shared memory.

- For instance, in a multi-core system,
  - each thread runs on a separate core.

Programming Task Parallelism

- Example and Exercise:
  - Implement the task-parallel computation of mean, median, and mode
    1. using threads in Java
    2. using P-Threads in C/C++
    3. using OpenMP in C/C++

Request Parallelism

Request Parallelism

- Requirement:
  - Scalable execution of repetitive but independent tasks in parallel, with dynamic arrival
- Solution:
  - As independent requests (for services) arrive,
  - Each request is assigned to a task in parallel
    - while other such tasks are servicing previous requests
- (Natural) Systems Fit:
  - Client-Server Model
- Examples:
  - E-mail Server, Web-Server, Cloud

Request Parallelism – Implementation and Performance

- This is typically implemented as a multi-threaded server:
  - A pool of threads are maintained
  - Each new request is assigned to a free thread
  - On completion of (servicing the assigned) request,
    - the thread de-allocates any resources previously allocated and
    - is marked free
- Performance Considerations:
  - Throughput
    - Number of requests serviced per unit time
  - Response Time
    - Turn-around time per request

ML Algorithms – Training Phase vs. Inference Phase

- During the inference phase or the prediction phase:
  - Request parallelism may be used to
    - deploy the model and
    - provide efficient inferences or predictions
- For the individual user submitting a request:
  - Response time is important (even if not critical)
    - Thumb rule in the practical world:
      - Any request on the Internet, the Web, or the Cloud must be serviced within 3 seconds
    - e.g. Recommender System on an e-commerce site
- For the provider:
  - Throughput is business-critical, while the
  - Average response time is important

Parallelization of ML Algorithms - Examples

Ensemble Methods

- Multiple ML algorithms (or learners) are trained on the same dataset:
  - The combination (ensemble learner) is expected to perform better than any of the individual learners.
- e.g. Bagging and Boosting

Bagging or Bootstrap Aggregation

- Generate m bootstrap data sets D_1, D_2, … D_m from the given data set:
  - Bootstrapping is the selection of random points with replacement
- Train each of the new data sets D_j to fit a model M_j and
  - Combine them
    - e.g. by taking the majority output of all classifiers or average of all the regressors.
- Bagging can be easily task-parallelized:
  - Tasks T_1, T_2, … T_m can each run as a different thread (i.e. on a different processor):
    - Where each T_j consisting of bootstrapping from the given set and training to obtain a model M_j

Bagging: Parallelization

- The parallelization discussed (see last slide) is for the training phase.
  - The inference phase requires a combination to be implemented.
    - This is easy (e.g., majority voting or averaging) and
    - parallelization is not critical
      - because m is not large (compared to n, the size of the dataset)

- Note:
  - Typically, bootstrap size B_{size} (or sample size) may be large
  - In bagging:
    - $B_{size} = n$ but
    - The number of bootstraps, m, is small.

AdaBoost (or Adaptive Boosting)

- Boosting:
  - Multiple learners $y_j(x)$ are trained on a weighted form of the training set.
  - Weights for each learner $y_j(x)$ are obtained from the performance of the previous learner $y_{j-1}(x)$
    - For instance, points that are misclassified by previous classifiers
      - receive greater weights in subsequent classifier(s)

AdaBoost (or Adaptive Boosting) [contd.]

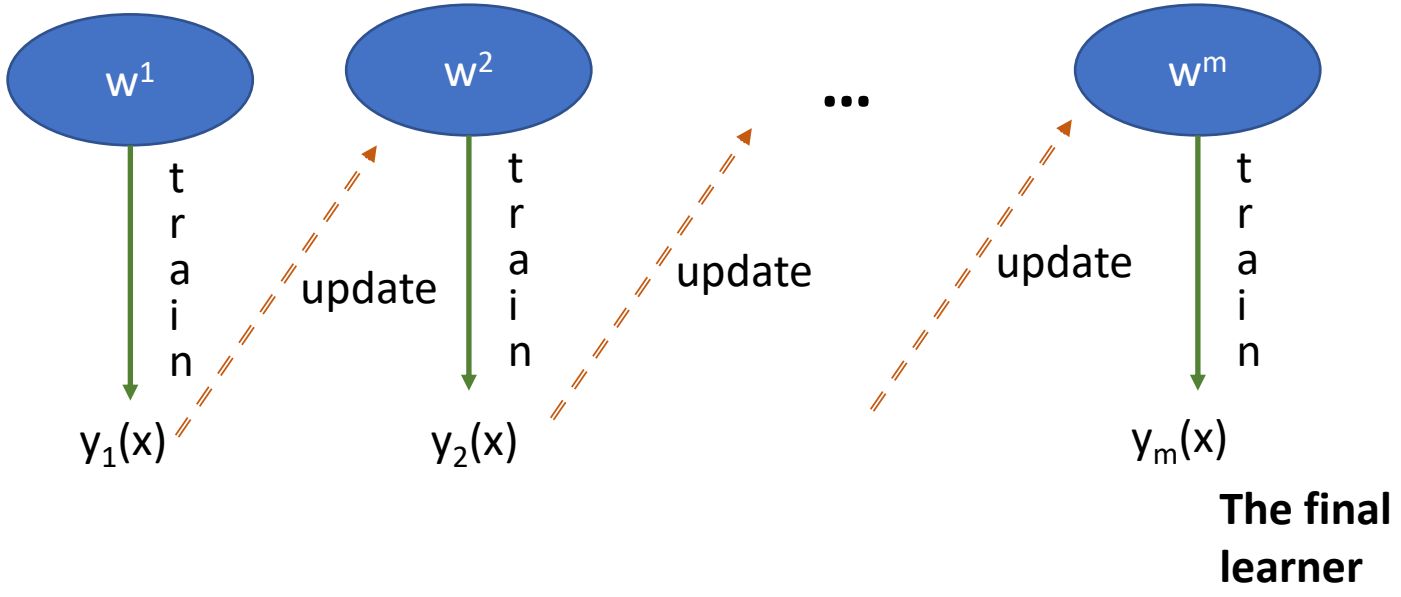
1. Initialize the data weighting coefficients $W^1 = 1/N$ for n= 1..N
2. for j = 1 to m
    1. Train a learner $y_j(x)$
    2. evaluate error e_j
    3. Update weighting coefficients $W^{j+1} = f(W^j, y_j, e^j)$
3. Make predictions using the final model y_m

AdaBoost (or Adaptive Boosting) [contd.]

1. Initialize the data weighting coefficients $W^1 = 1/N$ for n= 1..N
2. for j = 1 to m
   1. Train a learner $y_j(x)$
   2. evaluate error e_j
   3. Update weighting coefficients $W^{j+1} = f(W^j, y_j, e^j)$
3. Make predictions using the final model y_m

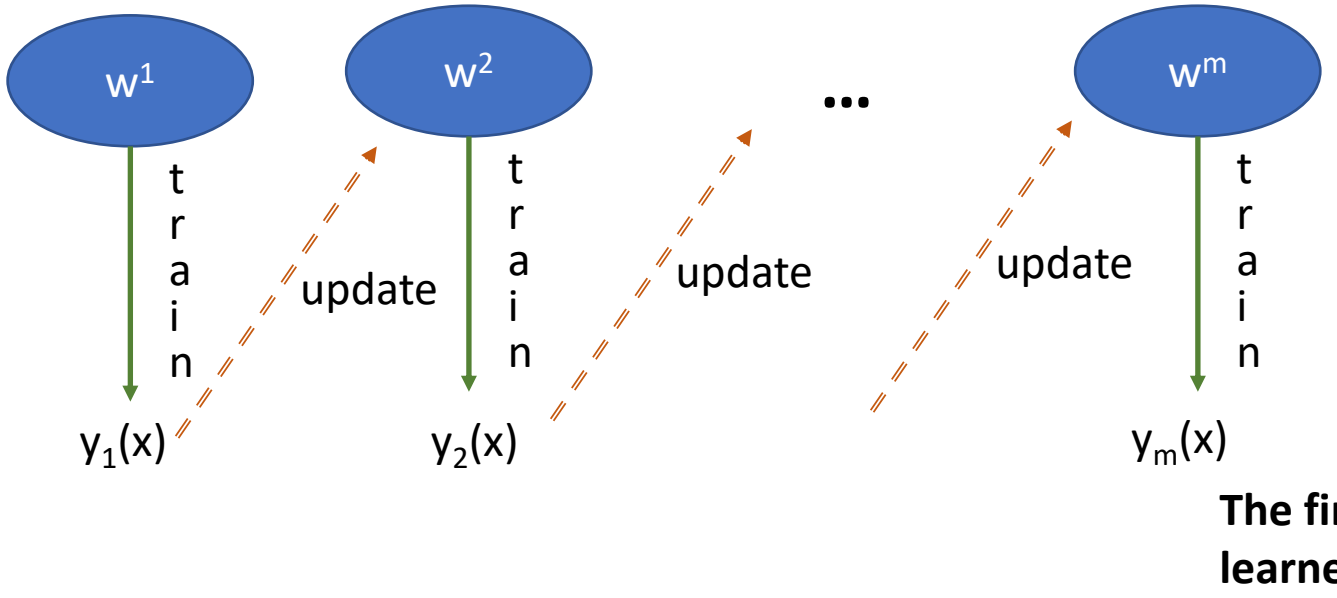
There is a sequential dependency!

This is not easily parallelizable!



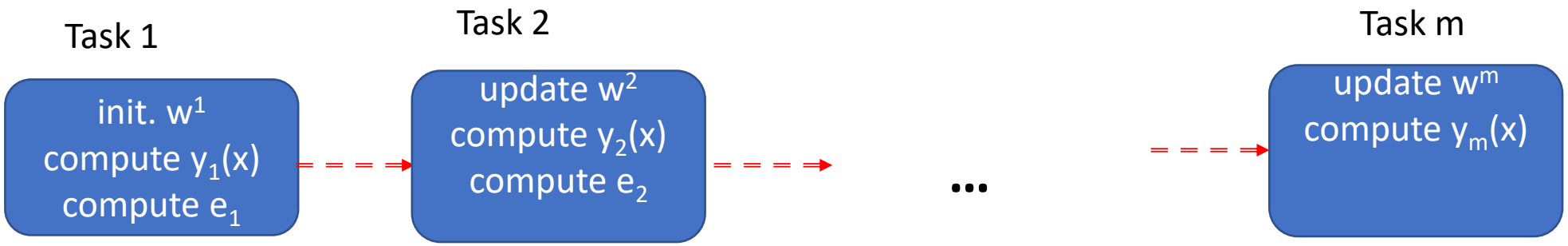
AdaBoost: Pipelining?

1. Initialize the data weighting coefficients $W^1 = 1/N$ for n= 1..N
2. for j = 1 to m
   1. Train a learner $y_j(x)$
   2. evaluate error e_j
   3. Update weighting coefficients $W^{j+1} = f(W^j, y_j, e^j)$
3. Make predictions using the final model y_m



While this algorithm is not amenable for data parallelism or for task parallelism, software pipelining may be attempted!

AdaBoost: Software-Pipelined



This pipeline provides speedup(m) > 1

only if computation of y_j and e_j can proceed in parallel with update w^{j+1}

BITS Pilani

WILP

AIML CLZG516

ML System Optimization

Shan Sundar Balasubramaniam

innovate achieve lead

BITS Pilani

WILP

AIML CLZG516
ML System Optimization

Session 4: 11 Jun. 2023

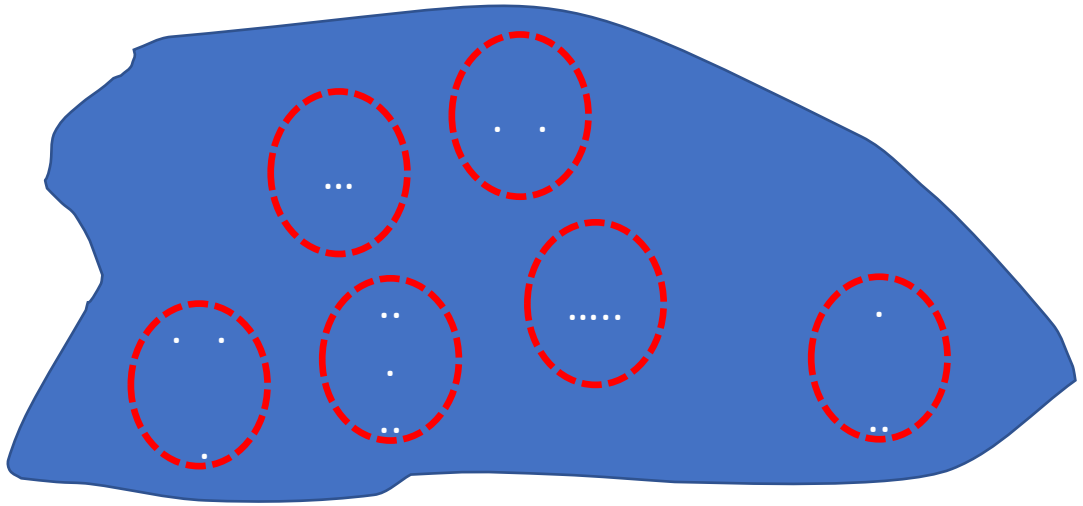
Parallelization of ML Algorithms

- Example: k-Means
- Issues with Large Data Size

Scale-out Clusters – Distributed Memory Programming

Example: Data Clustering using k-Means

- Data Clustering is a classic data analytics problem:
  - Given a set of data points group them into disjoint subsets – clusters – such that:
    - Each cluster is cohesive
    - Different clusters are well-separated



Points are in Euclidean space

K – means Clustering

Inputs: Dataset D, A positive integer k

Output: A partition Cs of D with size k

(i.e., k disjoint clusters covering all points in D)

Approach:

1. Chose k data points (as representatives) from D, say c_1, c_2, … c_k
2. Assign each point x in D to the cluster C_j ;
   whose that has the closest center c_j
3. Choose k new representatives based on
   minimizing local average distance within each cluster [Notion of ***cohesion***]
4. Iterate steps 2 and 3 until (the cluster centers converge)

K – means Clustering using map-reduce

- Step 1: "select representative points" for clusters $C_j = \{ c_j \}$ for j=1 to k
- Step 2:
  - map "compute distance" on D x Cs where Cs is the set of clusters
  - map "assign point to the closest cluster" on D
    - This requires: reduce min on point-cluster distances
- Step 3: for each cluster C_j compute its centroid (i.e., mean)
  - map on Cs:
    - $c_j = (\text{reduce} + \ C_j \ )/|C_j|$
- Repeat Steps 2 and 3 until all c_j converge

$D \times Cs = \{ (x, c_j) \ldots \}$
map comp_dist D x Cs

K – means Clustering using map-reduce

- Step 1: "select representative points" for clusters
- Step 2:
  - map "compute distance" on D x Cs where Cs is the set of clusters
  - map "assign point to the closest cluster" on D
    - This requires: reduce min on point-cluster distances
- Step 3: for each cluster C_j compute its centroid (i.e., mean)
  - map on Cs:
    - $c_j = (\text{reduce} + C_j)/|C_j|$
- Repeat Steps 2 and 3 until all c_j converge

{ (xi, d1j) }
= map comp_dist D
xi is a point in D
dij = distances of xi to clusters

reduce min dij

This reduce is required to return the cluster (with the min distance) and not the min distance:

Refer to reduce-key vs. reduce-val in Spark!

K – means Clustering

- Exercise: Implement k-means clustering using map and reduce.
- [Hints:
  - Step 1: "select k representative points" for clusters (randomly)
  - Step 2
    - map "compute distance" on D x Cs where Cs is the set of clusters
    - map "assign point to the closest cluster" on D
      - This requires: reduce min on point-cluster distances
  - Step 3b: compute the centroid (i.e., mean of) C_j
    - c_j = (reduce + C_j)/$|C_j|$

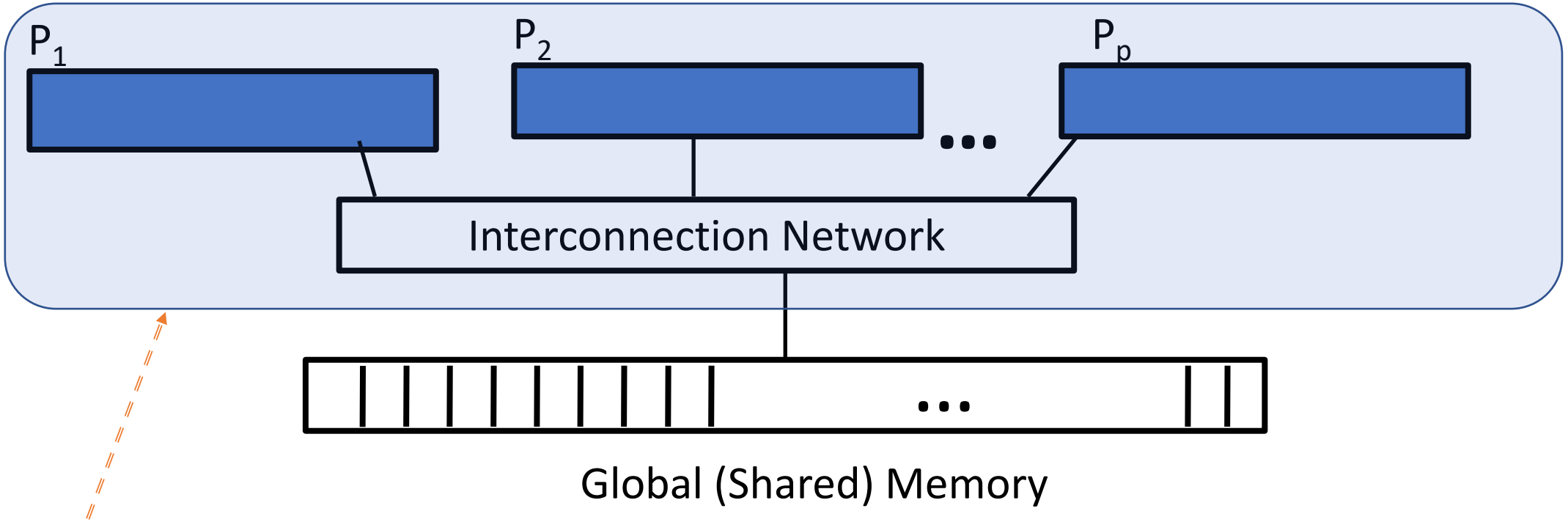
This follows a programming pattern named ***iterative map-reduce*** where map-reduce programming steps applied inside a loop.

Exercise: Speedup of k-means using map-reduce

- For each of the steps:
  - Calculate the speedup (and the number of processors)
- T_{seq} = I*(|D|*(k+k)+(k*|C|)) [Step 2: k distances req. k steps;
- min. computation req k-1 steps;
  - I is number of iterations
- T_{par} (p) = I*(|D|/p*(k+k)+(|C|) - assuming step 3 is done with only k processors; |D|/p points per processor in step 2
- p processors; k<p
- Speedup (p) = T_{seq} / T_{par} = (|D|*2*k + k*|C|)/((|D|/p)*(k+k)+|C|)
- ~p (close to ideal)

Parallel Programming: *Shared Memory* Model

So far we have looked at a target environment that uses a shared memory model:



e.g. a multi-core chip

Multi-threaded Programming:

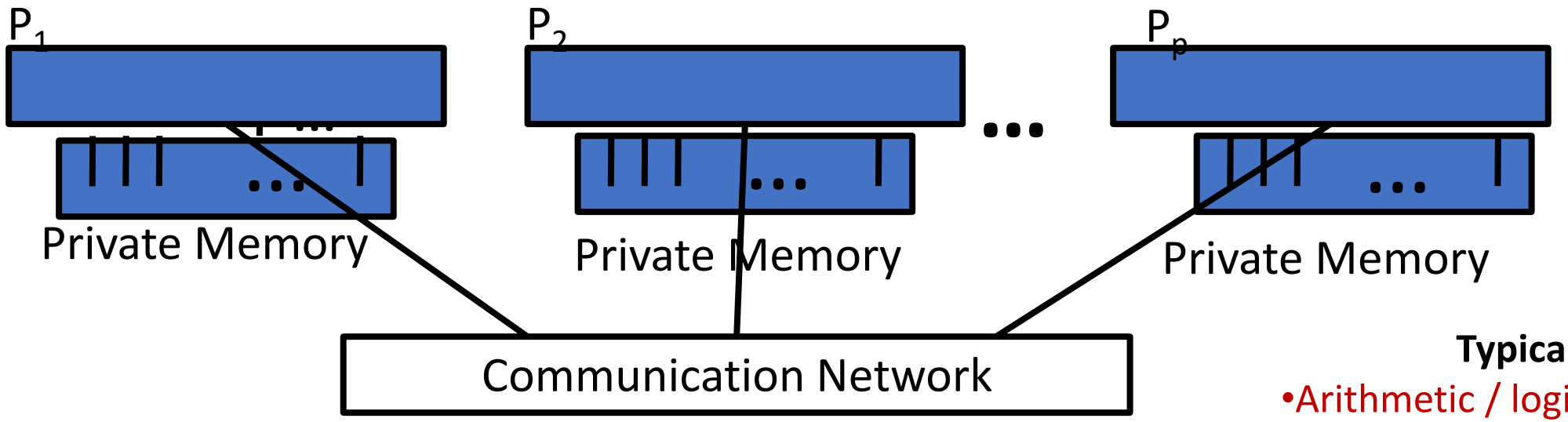
each thread runs on a separate core

Large volumes of Data

- When the volume of data that we have to process is in 100s of GB if not in TB,
  - Then all the data cannot be kept in one computer
  - And brought into memory for processing
- We a model where data can be stored on multiple computers (i.e., their hard disks)
  - All of which participate in computing.
- This leads us to a distributed computing model (aka message passing model)

Algorithm Design – Parallel: *Distributed Memory* Model

Target environment:



Typical Instructions

- Arithmetic / logic operations,
- Load / Store, and
- Jump / Branch
- **Send / receive**

Distributed Programming:

- a program is made of multiple processes
- *each process runs on a separate computer*
- *processes exchange messages (i.e., data for collaboration)*

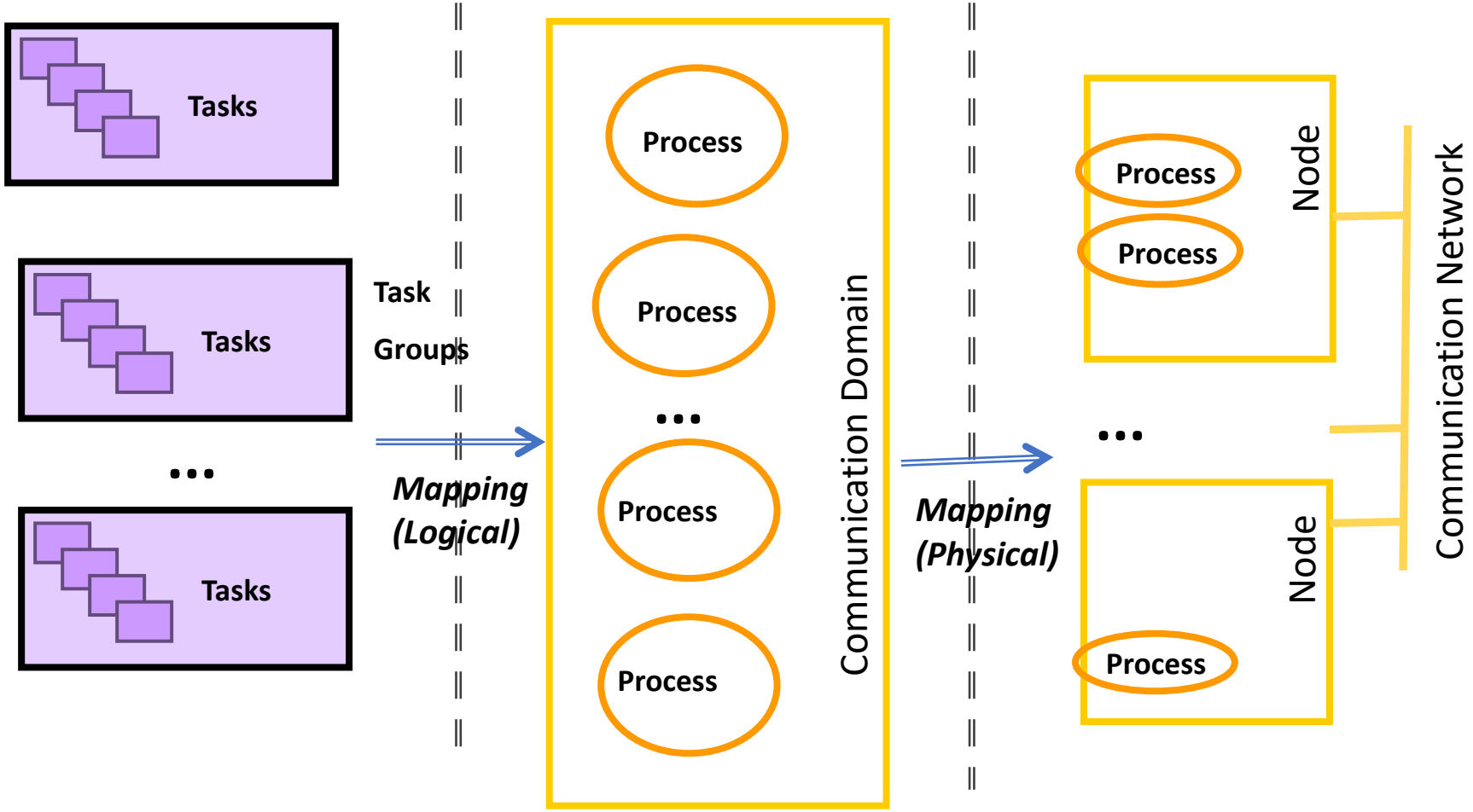
e.g. a cluster

Parallel / Distributed Computing

- A parallel or distributed program is made of multiple tasks that *collaborate* (to achieve a common outcome).
  - Collaboration is achieved by communication:
    - exchange data using shared memory
      - i.e. Task A writes to location L; Task B reads from location L
    - exchange data by passing messages
      - i.e. Task A sends a message to Task B; Task B receives the message from Task A

- Multiple processes each with its own address space:

E.g. processes run on nodes connected in a network : (i) each node runs its own OS and (ii) each process is allocated its own (logical) address space that is mapped onto the (physical) resources of that node

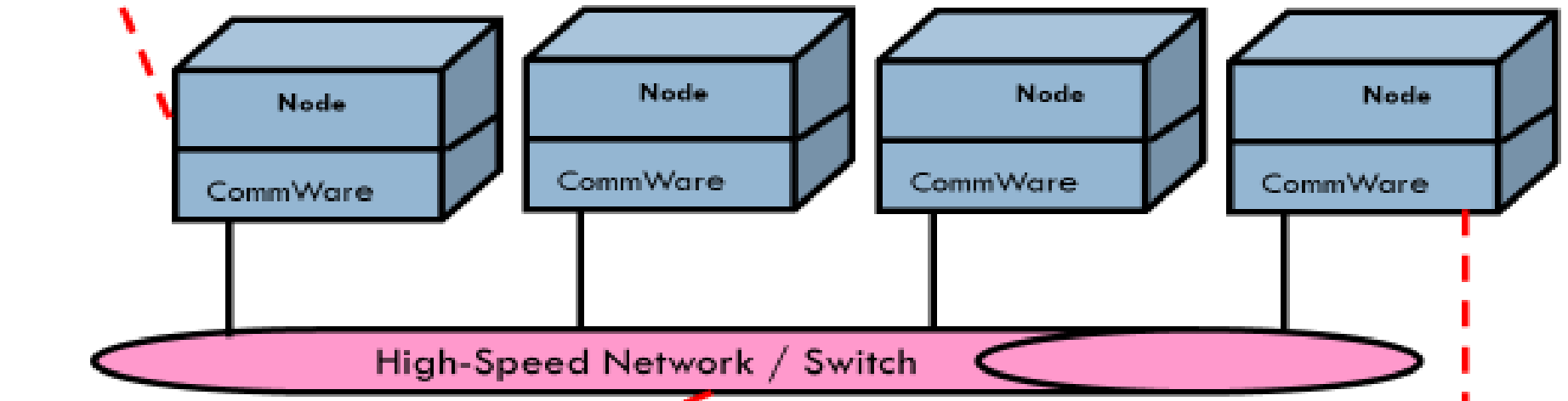


(Compute) Clusters

- Clusters are a modern example of distributed computing often used for high performance or big data computing.
  - Search engines - in the mid-90s - were using supercomputers at the back-end.
  - High obsolescence rate:
    - The supercomputers were getting obsolete (or unable to meet the computing needs) in five years or less.
  - Replacing a super-computer every five years was not cost-effective.
    - Clusters were available since 1980s and
    - Google realized that they are cost effective if older compute-nodes are replaced incrementally.
- Today, they are known as *scale out clusters* or *commodity clusters* and are used
  - with tens of thousands of nodes by Google, Facebook, Netflix, and others
  - for large scale processing of data.

Typical Cluster: Components - Base

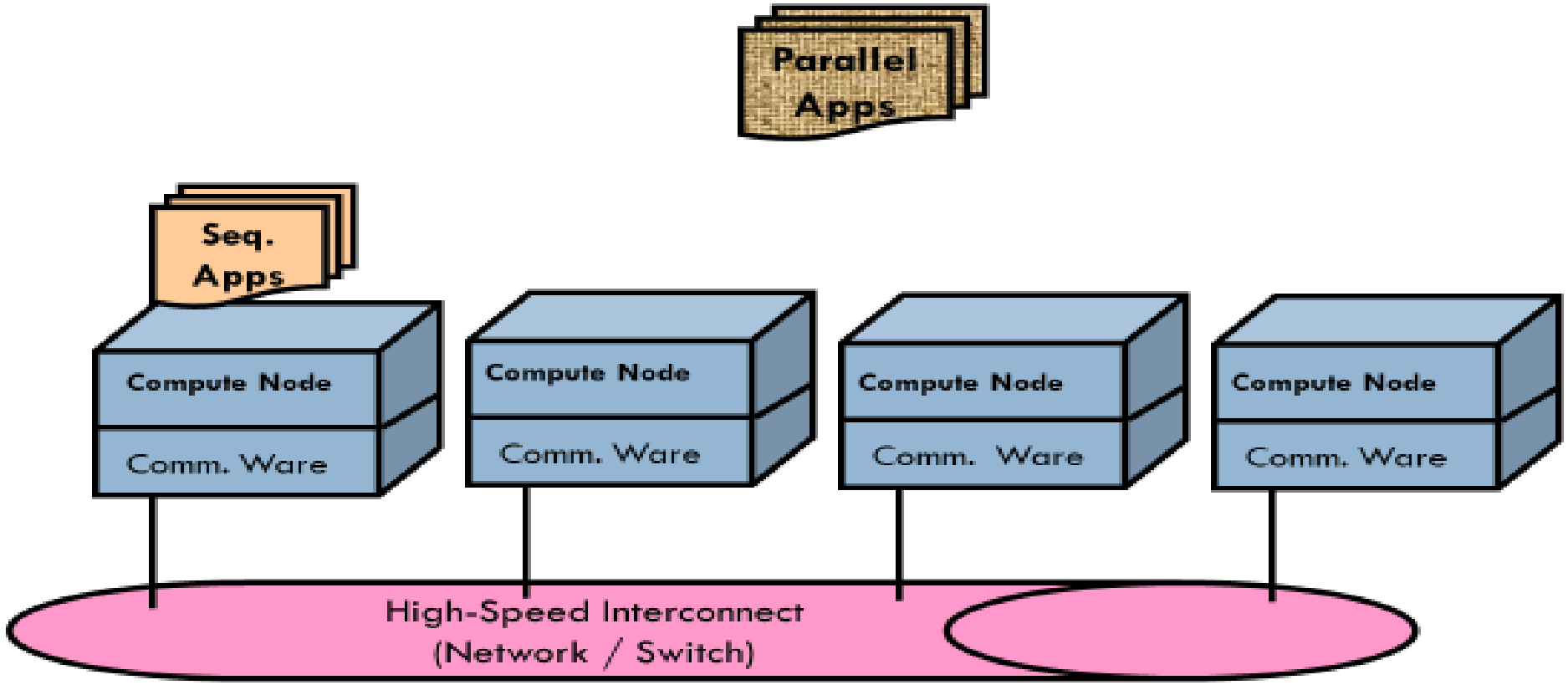
- **Processor+Memory+Storage**
- **OS+Run-time environment**



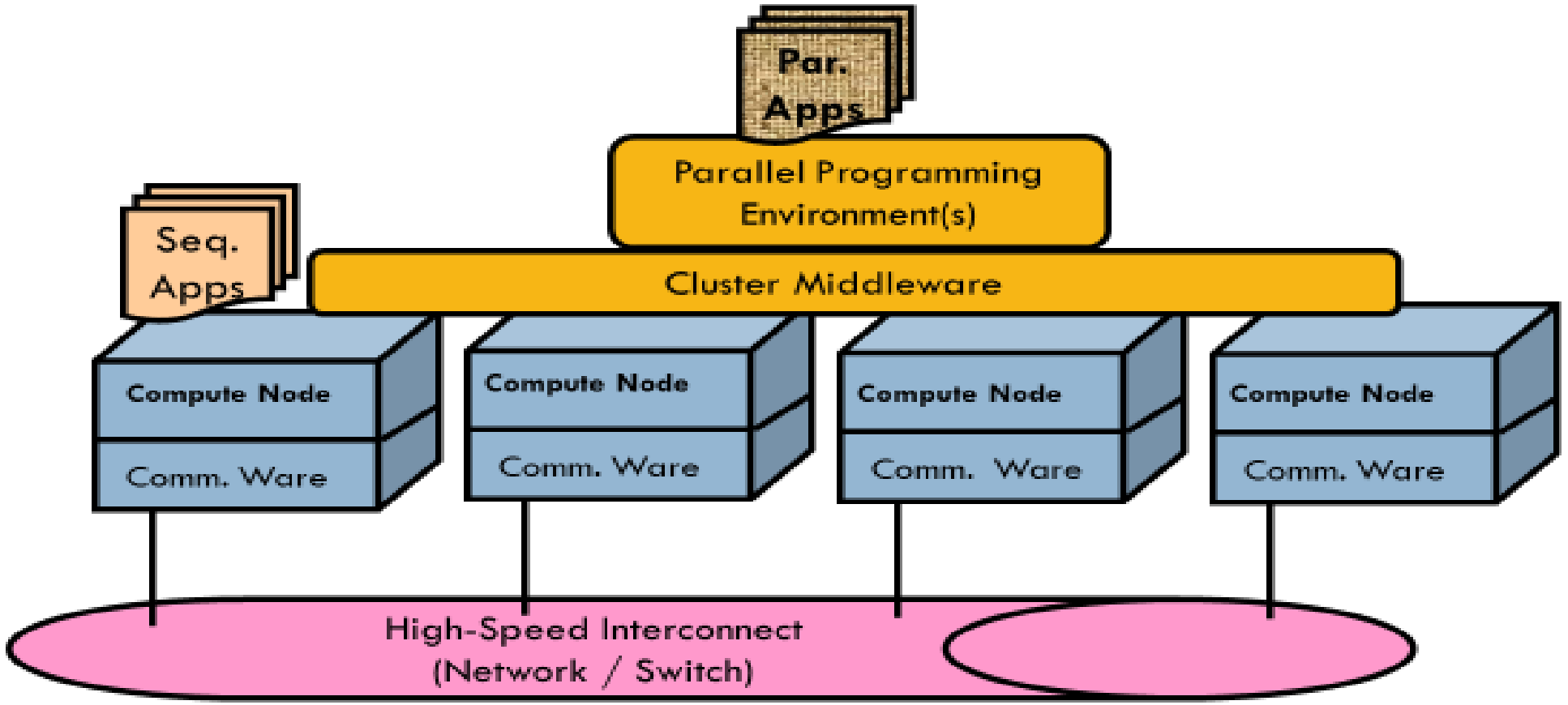
Gigabit Ethernet or Myrinet/Infiniband

- **Network Interface Cards**
- **Fast Communication Protocols and Services**

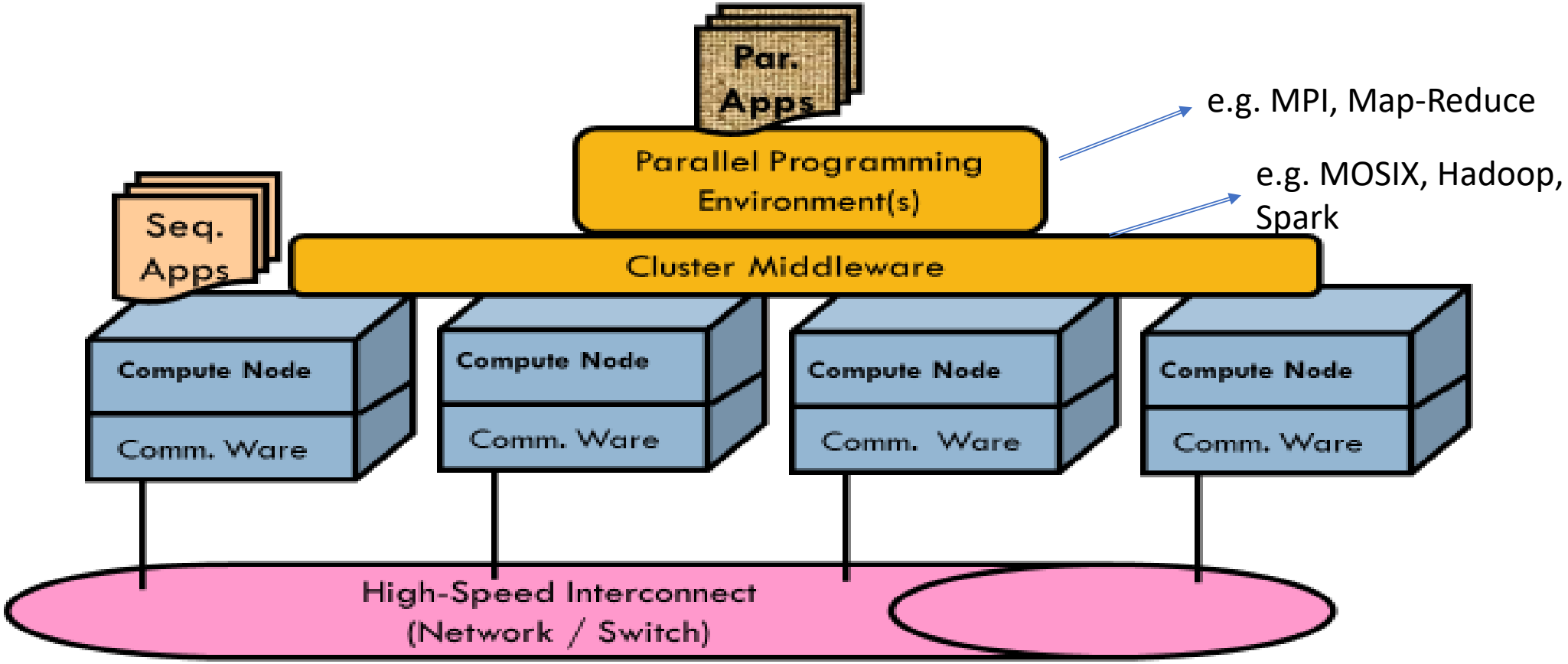
Typical Cluster - Requirements



Typical Cluster Architecture



Typical Cluster Architecture



Scale-out Clusters

- Case in-point:
  - Search engines - in the mid-90s - were using supercomputers at the back-end.
  - High obsolescence rate:
    - The supercomputers were getting obsolete (or unable to meet the computing needs) in five years or less.
  - Replacing a super-computer every five years was not cost-effective.
    - Clusters were available since 1980s and
    - Google realized that they are cost effective if older compute-nodes are replaced incrementally.
- Today, they are known as *scale out clusters* or *commodity clusters* and are used
  - with tens of thousands of nodes by Google, Facebook, Netflix, and others
  - for large scale processing of data.

Hadoop and Spark

- Apache Hadoop and Spark are platforms that run on clusters and support map-reduce programming
- Hadoop supports programming with data stored on files
- Spark supports in-memory distributed programming with RDDs (Resilient Distributed Data)
  - Programmable Data structures
    - Distributed in the memories of multiple nodes in a distributed system (e.g. a cluster)

Exercise

- Implement k-means on Spark
- Calculate – on paper - speedup of k-means using a cluster:
  - Calculate communication cost
- Understand:
  - the difference between this and the previous calculation (for shared memory programming)
- Questions:
  - How do you distribute the data initially?
  - Cost?
  - Pattern?

BITS Pilani

WILP

AIML CLZG516

ML System Optimization

Shan Sundar Balasubramaniam

innovate achieve lead

BITS Pilani

WILP

AIML CLZG516
ML System Optimization

Session 5: 18 Jun. 2023

Distributed Algorithms – Communication Overhead

Distributed ML Algorithms

- Example: k-Means

- Model Parallelism

K – means Clustering

Inputs: Dataset D, A positive integer k

Output: A partition Cs of D with size k

(i.e., k disjoint clusters covering all points in D)

Approach:

1. Chose k data points (as representatives) from D, say c_1, c_2, … c_k
2. Assign each point x in D to the cluster C_j ; 
   that has the closest center c_j
3. Choose k new representatives based on
   minimizing local average distance within each cluster [Notion of ***cohesion***]
4. Iterate steps 2 and 3 until (the cluster centers converge)

K – means Clustering using map-reduce

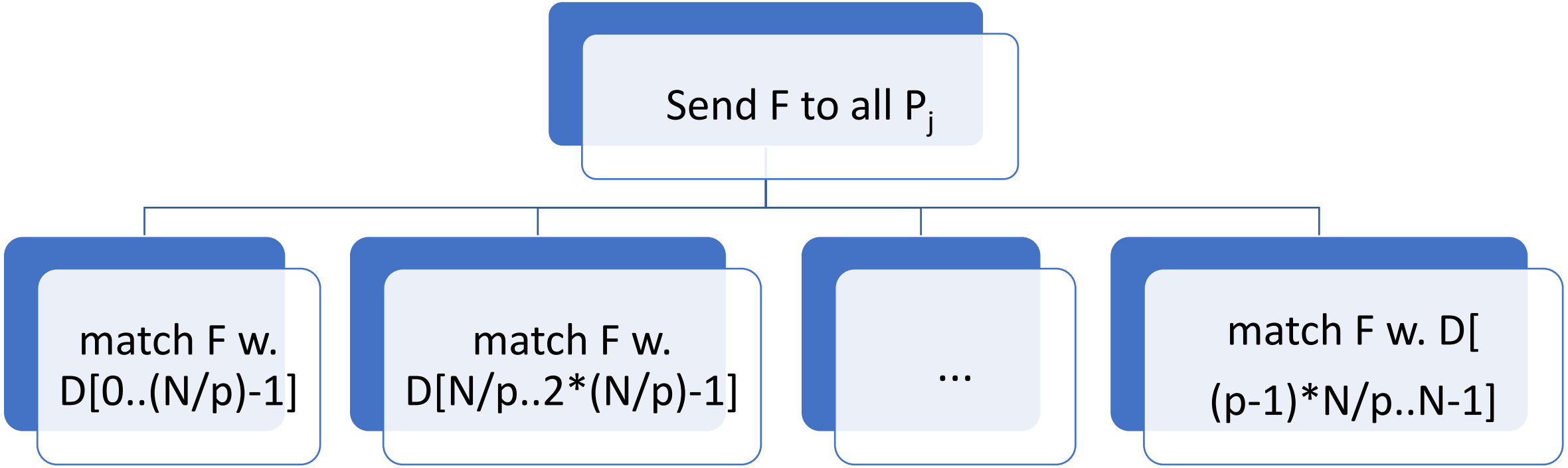
- Step 1: "select representative points" for clusters $C_j = \{ c_j \}$ for j=1 to k
- Step 2:
  - map "compute distance" on D x Cs where Cs is the set of clusters
  - map "assign point to the closest cluster" on D
    - This requires: reduce min on point-cluster distances
- Step 3: for each cluster C_j compute its centroid (i.e., mean)
  - map on Cs:
    - $c_j = (\text{reduce} + \ C_j \ )/|C_j|$
- Repeat Steps 2 and 3 until all c_j converge

$D \times Cs = \{ (x, c_j) \ldots \}$
map comp_dist D x Cs

Data Parallel Execution - Examples

Fingerprint Matching:

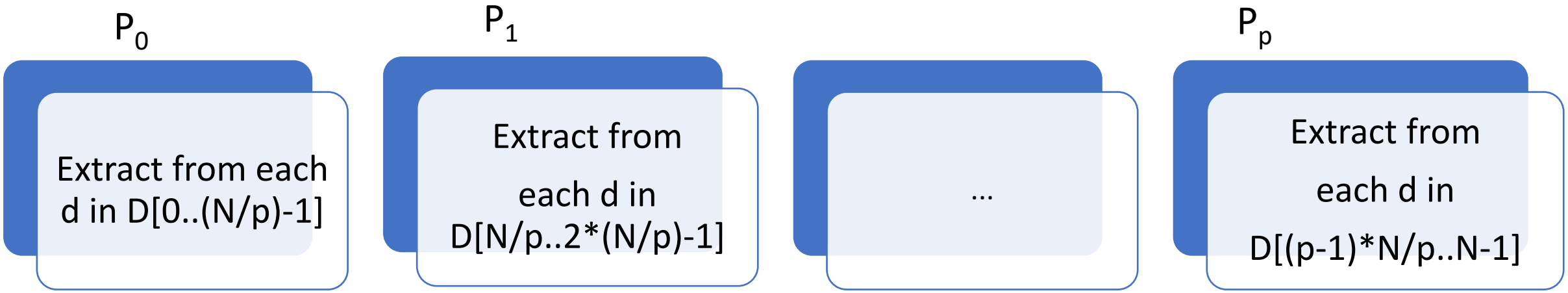
- Match a given print F with a database D of prints available;
- Assume D is distributed.



Communication cost?

Data Parallel Execution - Examples

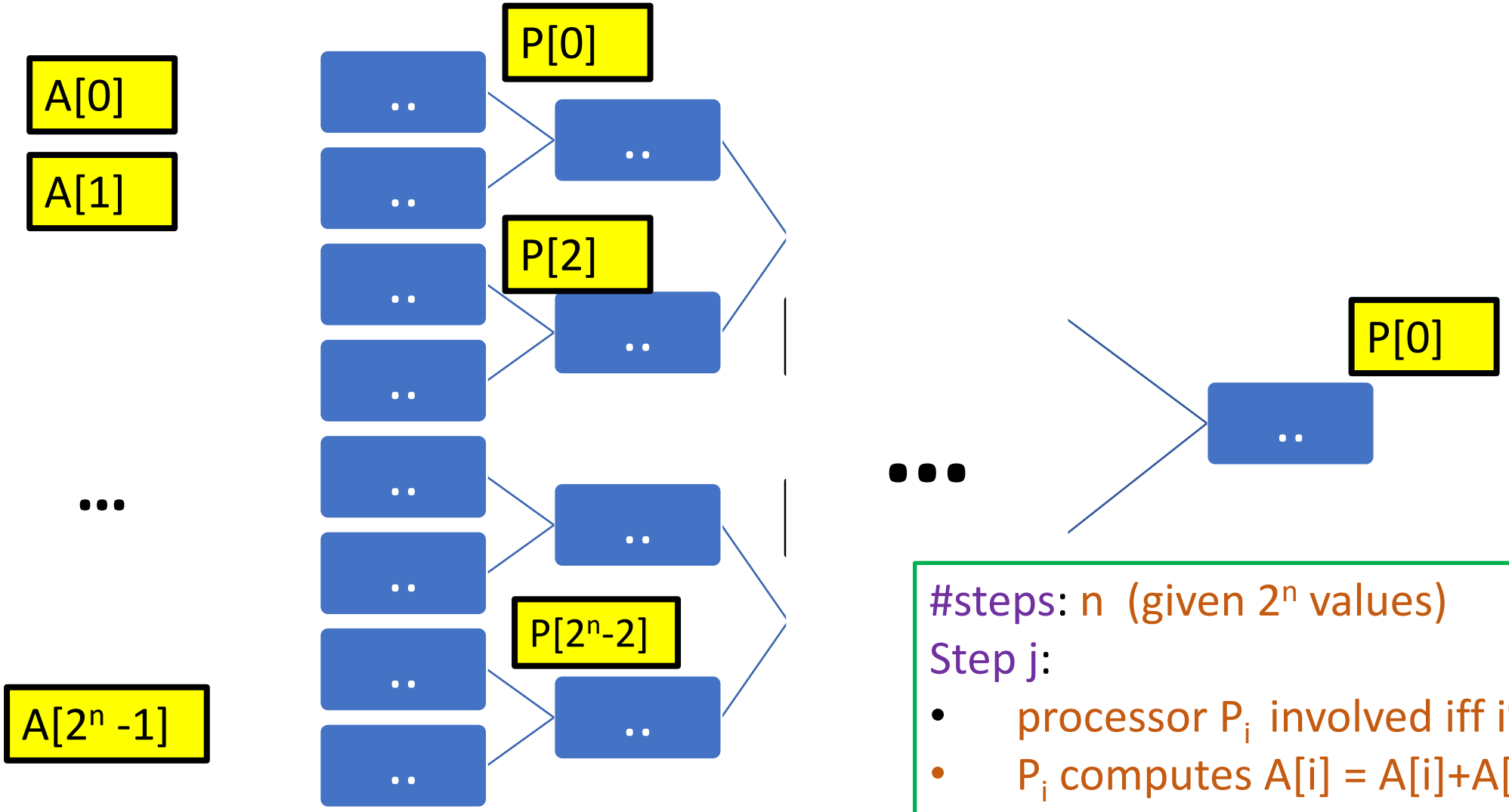
Extracting keywords from each document in a distributed collection D:
[Assume|D| = N]



P_0 : Collect keywords from all processors

Communication cost?

Example: Parallel Summation (Shared memory)



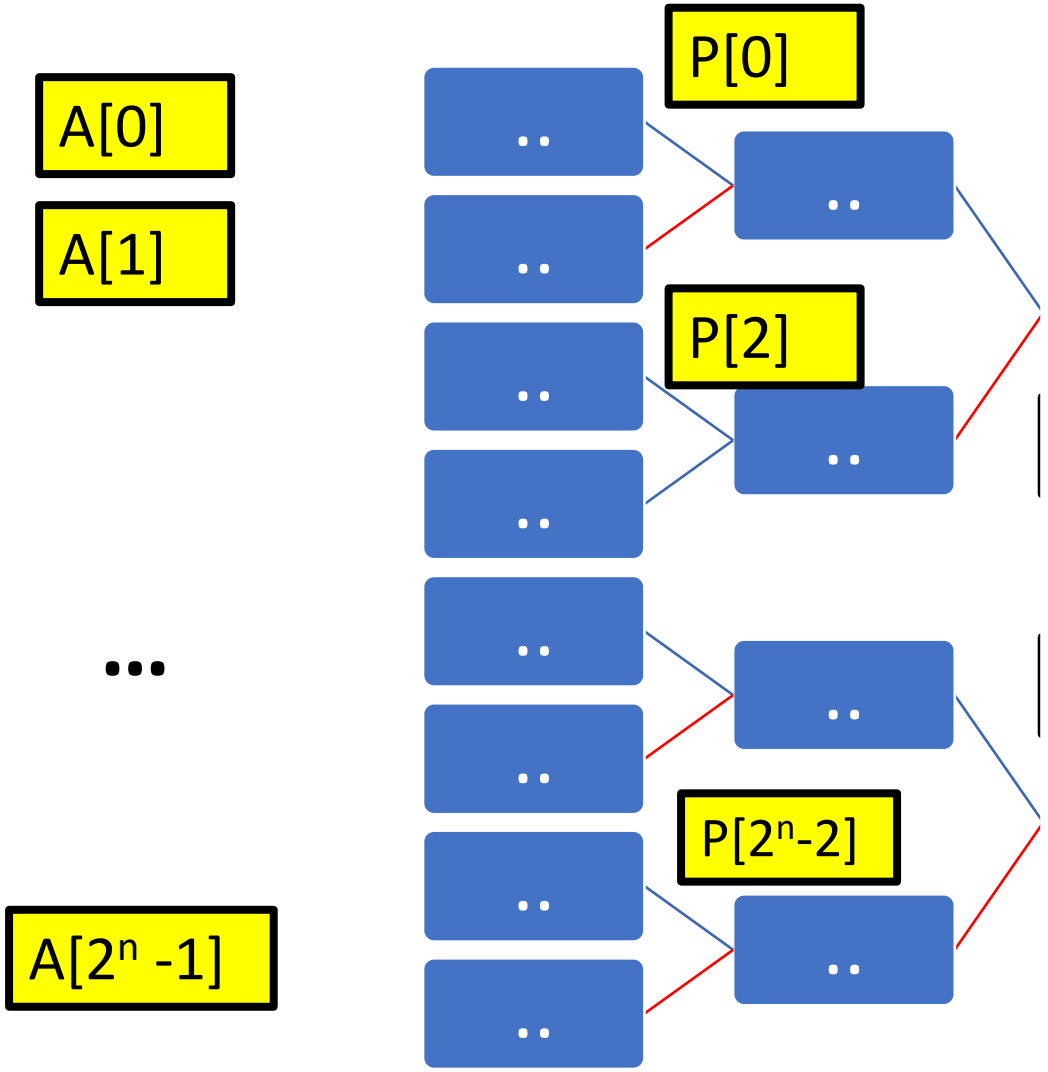
#steps: n (given 2^n values)

Step j:

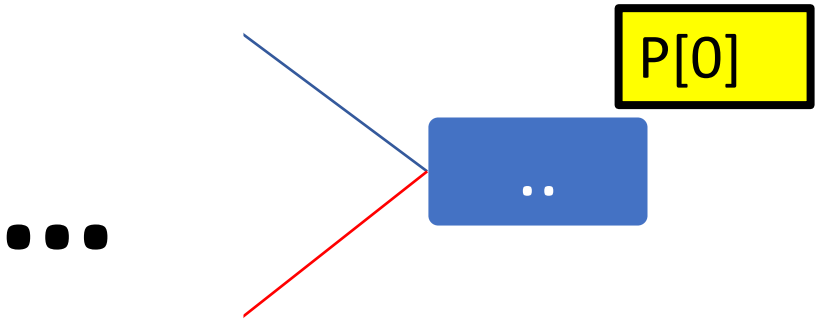
- processor P_i involved iff $i\%2^j == 0$
- P_i computes $A[i] = A[i]+A[i+2^{j-1}]$

A is in (global) shared memory

Example: Parallel Summation (Distributed memory)



indicates communication



#steps: n (given 2^n values)

Step j: (P_i has A[i])

- processor P_i involved iff $i\%2^j == 0$
- $A[i+2^{j-1}]$ is sent to P[i]
- P_i computes $A[i] = A[i]+A[i+2^{j-1}]$

Algorithm Design – Speedup: Caveat

- Summation by reduction executed on a distributed system:
  - before processor P_i performs $A[i] = A[i] + A[i+2^{j-1}]$
    - there is communication involved: $A[i+2^{j-1}]$ sent from $P_{i+2^{j-1}}$ to P_i
- The time complexity of a reduction algorithm will increase
  - i.e. *speedup will decrease*.

- Speedup (of summation by reduction of **N** values with **N/2** processors):

$$S(N,p) = T_{seq}(N) / T_{par}(N,N/2)$$
$$= ((N-1)*T_{add}) / (\log N*(T_{add} + T_{msg}))$$

where T_{add} is the time taken for adding two values
and T_{msg} is the time taken for sending (and receiving) a message

Algorithm Design – Speedup: Caveat

- Speedup (of summation by reduction of **N** values with **N/2** processors):

  $S(N,N/2) = T_{seq}(N) / T_{par}(N,N/2)$

  $\approx (N/\log N) * (1 / (1 + T_{msg}/T_{add}))$

- T_{msg} includes
  - set-up cost (for *send* and *recv*), which is typically fixed and
  - transmission cost (which depends on the length of the message)
- Typically, $T_{msg} >> T_{add}$

	1990s	Today (2020s)
T_{msg}	*ms*	*us*
T_{add}	*us* (or 10s of *ns*)	< 1*ns*

ms milli-seconds
us micro-seconds
ns nano-seconds

- Implication: Speedup << N/logN

Reduce as in Google's Map-Reduce

- Pragmatics of ***reduce*** (*similar to that of* **map**):
  - (Assumption) Input Data partitioned and stored
  - Management of Messaging / Communication
  - Spawning of processes, Scheduling, and Load Balancing
  - Recovery from process / node failures

Communication is managed transparently:

i.e., programming is easier,

but communication cost is still the same!

K – means Clustering: Algorithm outline

Inputs: Dataset D, A positive integer k

Output: A partition Cs of D with size k

(i.e., k disjoint clusters covering all points in D)

Approach:

1. Chose k data points (as representatives) from D, say c_1, c_2, … c_k
2. Assign each point x in D to the cluster C_j :
   that has the closest center c_j
3. Choose k new representatives based on
   minimizing local average distance within each cluster [Notion of ***cohesion***]
4. Iterate steps 2 and 3 until (the cluster centers converge)

K – means Clustering on a distributed system

- Assume D is distributed in **p** processors (or nodes)
  - Each processor p_i has D_i of size (|D|/p points)
- Step 1: "select representative points" for clusters $C_j = \{ c_j \}$ for j=1 to k
- Step 2:
  - Communicate c_j (for j = 1 to k) to all p nodes (i.e. processors)
  - On each processor p_i
    - for each point x in D_i
    - {
    - for j=1 to k { d_j = distance(x,c_j); }
    - assign x to the cluster with minimum d_j
    - } // each p_i has k clusters

...

K – means Clustering on a distributed system

- Assume D is distributed in **p** processors
  - Each processor p_i has D_i of size (|D|/p points)
- Step 1: "select representative points" for clusters $C_j = \{ c_j \}$ for j=1 to k
- Step 2:
  - Communicate c_j (for j = 1 to k) to all p nodes (i.e. processors)
  - On each processor p_i, "compute distances":
    - for each point x in D_i
    - {
    - for j=1 to k { d_j = distance(x,c_j); }
    - assign x to the cluster with minimum d_j
    - } // each p_i has k clusters
- Step 3: collect and distribute clusters to k different processors
  - for each cluster C_j in processor p_j
    - c_j = (reduce + C_j)/|C_j|
- Repeat Steps 2 and 3 until all c_j converge

K – means Clustering on a distributed system

- Assume D is distributed in **p** processors
  - Each processor p_i has D_i of size (|D|/p points)
- Step 1: "select representative points" for clusters $C_j = \{ c_j \}$ for j=1 to k
- Step 2:
  - Communicate c_j (for j = 1 to k) to all p nodes (i.e. processors)
  - On each processor p_i ⟸ map *(distributed)*
    - for each point x in D_i
    - {
    - for j=1 to k { d_j = distance(x,c_j); } ⟸ map *(shared mem.)*
    - assign x to the cluster with minimum d_j ⟸ reduce *(shared mem.)*
    - } // each p_i has k clusters
- Step 3: collect and distribute the clusters to **k** different processors
  - for each cluster C_j in processor p_j (j=1 to k) ⟸ map *(distributed)*
    - c_j = (reduce + C_j)/|C_j| ⟸ reduce *(shared mem.)*
- Repeat Steps 2 and 3 until all c_j converge

K – means on a distributed system: Communication cost

- Assume D is distributed in **p** processors
  - Each processor p_i has D_i of size (|D|/p points)
- Step 1: "select representative points" for clusters $C_j = \{ c_j \}$ for j=1 to k
- Step 2:
  - Communicate c_j (for j = 1 to k) to all p nodes (i.e. processors) ← k values
  - On each processor p_j, "compute distances":
    - for each point x in D_i
    - {
    - for j=1 to k { d_j = distance(x,c_j); }
    - assign x to the cluster with minimum d_j
    - } // each p_i has k clusters
- Step 3: collect and distribute clusters to k different processors ← |D| values
  - for each cluster C_j in processor p_j
    - c_j = (reduce + C_j)/|C_j|
- Repeat Steps 2 and 3 until all c_j converge

K – means on a distributed system: Implementation

- Use MPI (Message Passing Interface) to program on a distributed system:
  - Defaults to a SPMD model (i.e. same program/code on multiple processors but different data)
    - i.e. map is implicit
  - Constructs available for send, recv, reduce
- Alternative:
  - Use Hadoop map-reduce
  - Data stored on files (of nodes in a cluster)

K – means on a shared memory system: Implementation

- Use OpenMP library or P-Threads library) to program on a shared memory) system:
  - Defaults to a SPMD model (i.e. same program/code on multiple processors but different data)
    - i.e. map is implicit
  - Constructs available for send, recv, reduce
- Alternative:
  - Use multi-threaded programming in Java

K-means on a commodity cluster

- A cluster of nodes (workstations/servers)
  - Each node has a multi-core processor
- Use MPI programming on the cluster (in C or in Java)
  - The code for each node can be multi-threaded
    - Use OpenMP or P-Threads

K – means on Spark

- Distributed memory model (i.e. a cluster) where
  - Every node can be a shared memory processor (i.e. a multi-core processor)
  - Supports map and reduce as constructs
  - Programming in Java or Scala or other languages
  - In-memory processing
    - i.e. data may be stored on files,
    - but map and reduce work on data structures (RDDs) stored in memory
      - Unlike on Hadoop where map and reduce operate on files

Task Parallelism – Example

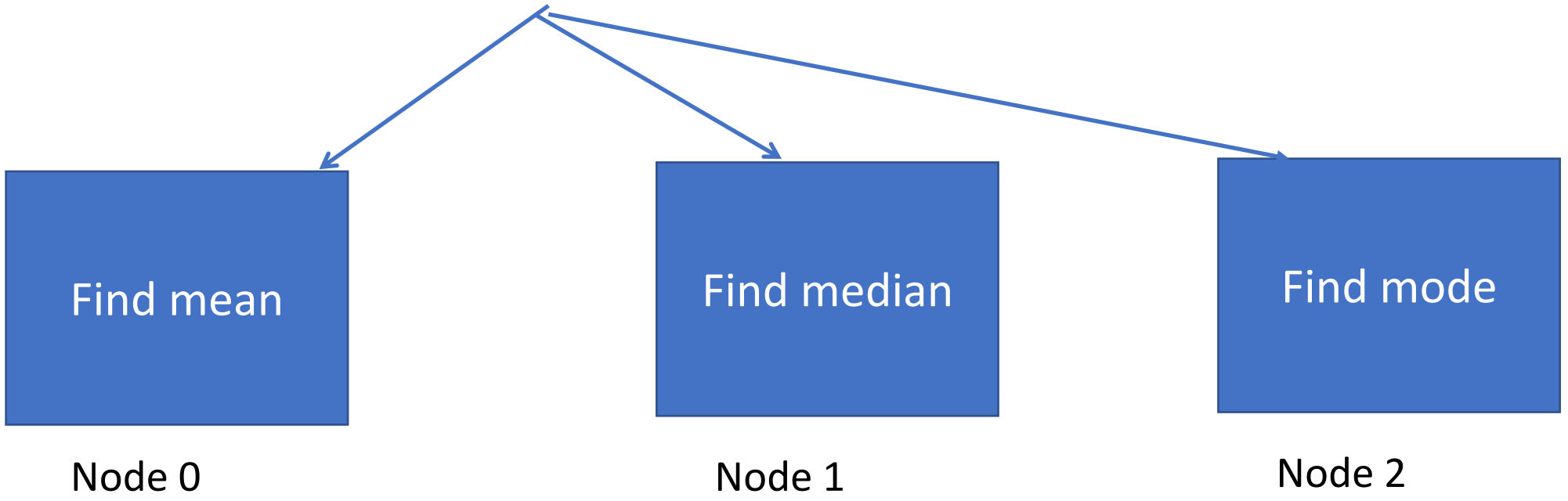
Recall Task Parallelism on Shared memory computers

The same is possible on Distributed memory systems!

Communication?

Assumption:

Input data is available on all nodes.

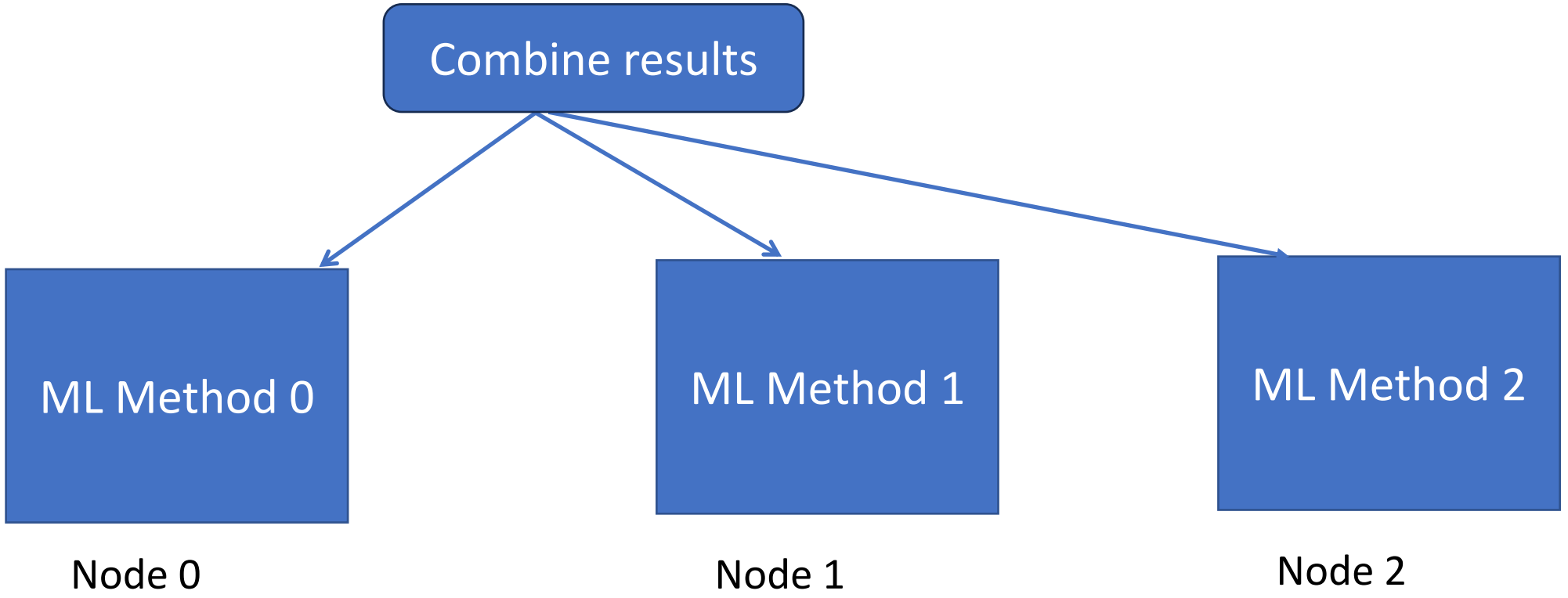


Ensemble Methods

- Task Parallelism – in the context of ML – is also known as Model parallelism
  - Recall the Bagging solution:
    - Multiple different models from the same data (via sampling)
  - Ensemble methods
    - Multiple different models from the same data (via different training methods)

Model Parallelism – Example: Ensembles

Assumption:
Input data is available on all nodes.



Communication Cost?

BITS Pilani

WILP

AIML CLZG516

ML System Optimization

Shan Sundar Balasubramaniam

innovate achieve lead

BITS Pilani

WILP

AIML CLZG516
ML System Optimization

Session 6: 25 Jun. 2023

Parallel/Distributed ML Algorithms

- Exercises: kNN, Decision Trees
- Assignment

Nearest Neighbors method for Classification

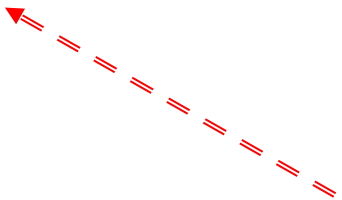
- Approach:
  - A (test) point **p** is assigned to a class **C**, to which belong a majority of the neighbors of **p**
    - This requires determining the nearest neighbors
  - The number of neighbors, K,
    - to be used to identify the class of a given test point is chosen externally
    - and the choice is non-trivial:
      - Too Large K ==> reduced accuracy
      - Too Small K ==> increased noise-sensitivity

K Nearest Neighbors: Algorithm

Inputs: Dataset D, positive integer k // D is (partially) labelled

Approach:

1. for each x in D:
    1. Compute distances to all other points in D
    2. Choose the k neighboring points nearest to x
    3. Identify the class C to which a majority of these neighbors belong
    4. x.class = C



Break ties arbitrarily!

K Nearest Neighbors: Algorithm: Analysis

Inputs: Dataset D, positive integer k // |D|=N

Approach:

1. for each x in D:
    1. Compute distances to all other points in D
    2. Choose the k neighboring points nearest to x
    3. Identify the class C to which a majority of these neighbors belong
    4. x.class = C

- $O(N^2)$ distance computations
- Each computation take O(d) time
    - (assuming d-dimensional Euclidean space).

- O(N) majority computations
- Each computation takes O(k) time

What about step 2?

K Nearest Neighbors: Algorithm: Analysis

Inputs: Dataset D, positive integer k

Approach:

1. for each x in D:
    1. Compute distances to all other points in D
    2. Choose the k neighboring points nearest to x
    3. Identify the class C to which a majority of these neighbors belong
    4. x.class = C

Step 2:

- Naïve approach: Sort the neighbors by distance! // O(NlogN) additional time

A popular alternative: Construct a KD-tree (using the distance metric)

- Combined cost for distance computations and tree construction:
    - O(d*N*logN), where for d-dimensional data

Shared Memory processor

Parallel Algorithm - Data parallel

- How do you parallelize kNN?
  - for each x in D:
  - Step 1: Distance Computation
    - N^2 processors: 1 distance computation each;
    - p processors: N^2/p computation in each processor
    - ==========
    - Step 2: Getting the nearest neighbors
  - Option 1: for each x in D:
    - Parallel-sort the neighbors by distance (and select the first k)
  - Option 2: for each processor P_j j = 1 to p
  - for each x in D_j
    - sort the neighbors (all points) by distance // sorting of |D| values
    - ==========
  - Step 3: for each processor P_j j = 1 to p
  - for each x in D_j
    - Finding the majority among k neighbors

Speedup: p

Speedup: p

Speedup: p

Online Algorithm – Request Parallelism

- Maintain a pool of threads
  - When a point p arrives, it is assigned to a (free) thread

- (thread == processor)

kNN: Distributed Algorithm

- How do you distribute kNN ?
  - Step 1: Distance Computation
  - Step 2: Getting the nearest neighbors
  - Step 3: Finding the majority

  - E.g. Step 1: point 1 (at node) – distance computation with all other points

    Communication cost? Entire dataset is broadcast?
    Distributed KD-Tree Construction?

    N points into p nodes; (N/p) points each

- Data set D has been divided into p nodes (N/p points each; N = |D|)

- Let point x be in node i
- Send point x to all p nodes
    - Parallel: In each node dist(x,y) is computed for all y in D_j j = 1 to p
    - Send back k nearest neighbors of x in node j to node I
    - node i has p*k potential neighbors; extract the k nearest
- O(k) per point

Decision Trees

- Approach:
  - Construct a tree where each node denotes a binary decision
    - Nodes in the tree correspond to features and the order of features is chosen based on the notion of **information gain** (IG)
    - Information gain is the entropy
      - entropy of the whole set
      - minus the entropy when a particular feature is chosen

Decision Tree Construction

- Algorithm ID3 (input dataset S)
  - If all examples have the same label
    - Return a leaf with that label
  - Else if there are no features left to test
    - Return a leaf with the most common label
  - Else choose the feature F that maximizes IG of dataset S as the next node
    - Add a branch from the node for each possible value f in F
    - For each branch:
      - Calculate S_f by removing F from the set of features
      - ID3(S_f)

Parallel/Distributed Tree Construction?

- When you branch assign each branch (corresponding to one value of a feature)
  - To a different task (task parallelism)
  - At each level : number of parallel tasks = number of possible values of a feature

Assignment

- Assignment will be released soon.
  - Assignment and Project will form a single end-to-end sequence of take-home team exercise involving:
    - Literature Survey,
    - Problem Formulation,
    - Design,
    - Implementation,
    - Testing and Demo
      - particularly for Performance
- Teams can include at most five students
  - Problems must focus on paralelization/distribution of ML algorithms
  - Target architectures can be multi-core, GPUs, or clusters or a combination thereof.
  - Any Programming language/environment or development platform may be used.

BITS Pilani

WILP

AIML CLZG516

ML System Optimization

Shan Sundar Balasubramaniam

innovate achieve lead

BITS Pilani

WILP

AIML CLZG516
ML System Optimization

Session 7: 9 July 2023

Distributed ML - Models and Platforms

- Implementation Issues
- The Parameter Server Model
- Stochastic Gradient Descent

ML problem and Error

- Given input dataset – a vector of size n,
  - Each training example $\mathbf{x}_i$ of **d** features (or dimensions) is associated
    - with a label $\mathbf{y}_i$ and
    - model parameters (likely corresponding to the features)
  - The problem is to predict **y** corresponding to an unseen example **x**
- Training error
  - Difference between the predicted y the actual label y' for an x

Model complexity

- Relation between model size (number of parameters) and data size (for training):
  - If there is too little data,
    - then a highly detailed model may overfit
  - If the model is too small,
    - then it may fail to capture relevant attributes
- Regularization addresses this issue

ML as regularized error minimization

- Training an ML model is minimizing the function F :
  - $F(w) = \Sigma_i L(x_i, y_i, w) + \Omega(w)$
    - where w denotes the set of parameters and
      - L is the loss function (i.e. prediction error) and
      - Ω is the regularizer that penalizes the model for complexity

Distributed ML

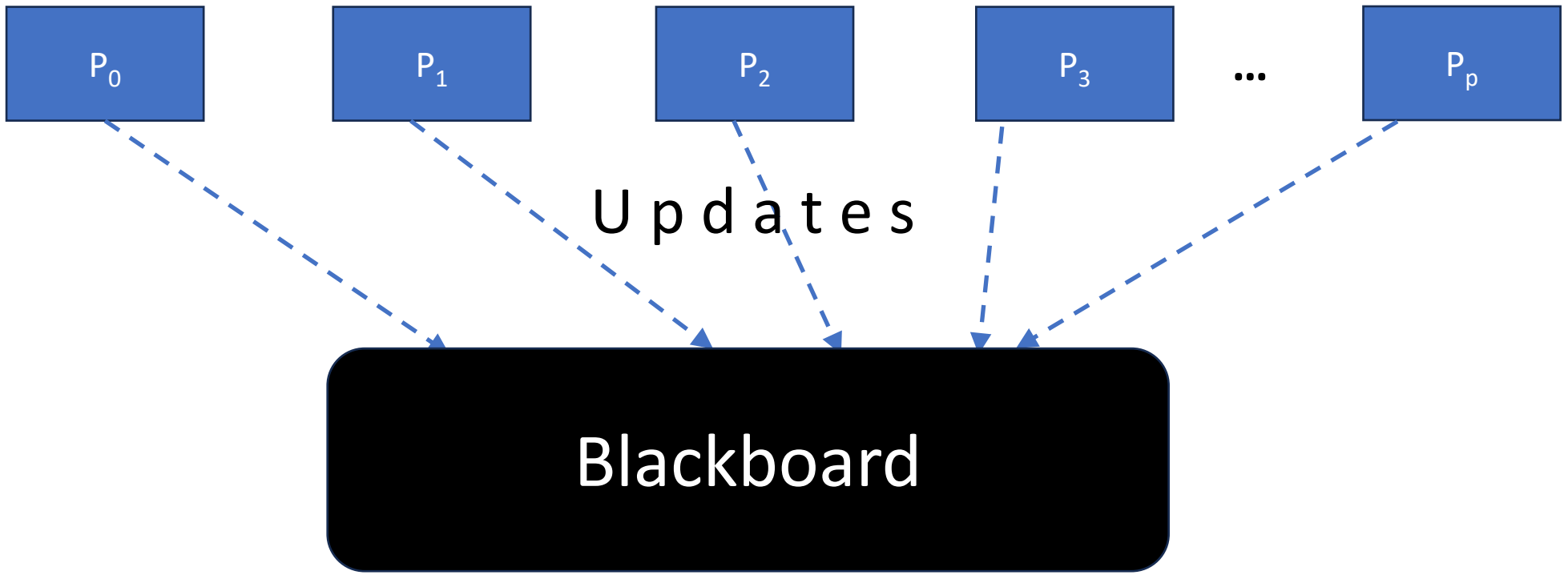
- Complexity:
  - Training Data size: from 1TB to 1PB
  - Model Size: 10^9 to 10^{12} parameters
- Examples:
  - Online Recommender System
    - millions of user profiles
  - Ad click predictor
    - each training example is a feature vector of high dimensions

Distributed ML – System Requirements

- In a distributed system,
  - the training data is partitioned among multiple nodes
    - and the nodes together learn the parameter vector w.
- The algorithm operates iteratively:
  - In each iteration,
    - every node independently uses its own training data to
      - Compute the changes to be made to **w** in order to move closer to an optimal value
    - Each node computes changes to **w** based only on its local data,
      - a central place is needed to aggregate these changes

Distributed Systems – Blackboard architecture

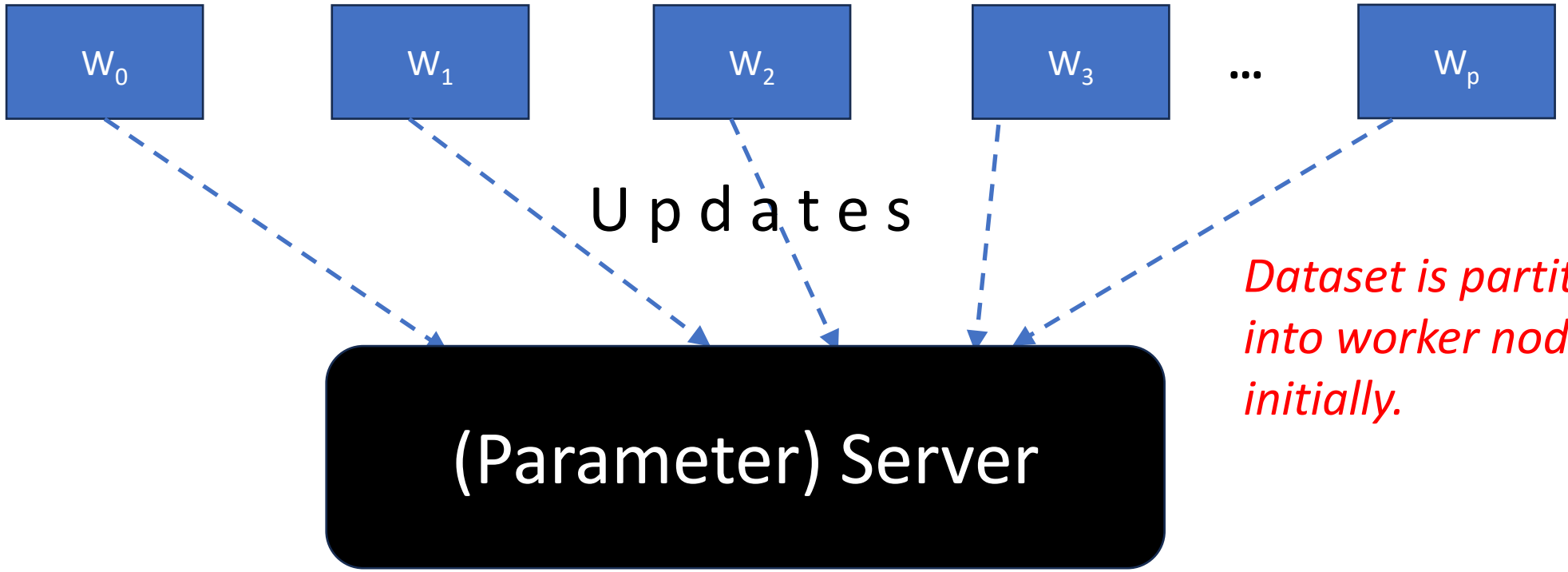
- Blackboard architecture is a pattern for distributed computation
  - where multiple nodes have to combine results
    - *computed locally, in parallel - see processes P_i below*



Regularized Error Minimization : A Distributed Architecture

In each iteration:

- Each worker node W_i pulls (current) parameters w from the server, computes F(w), and posts it back.
- Server updates and minimizes F(w)



Dataset is partitioned into worker nodes initially.

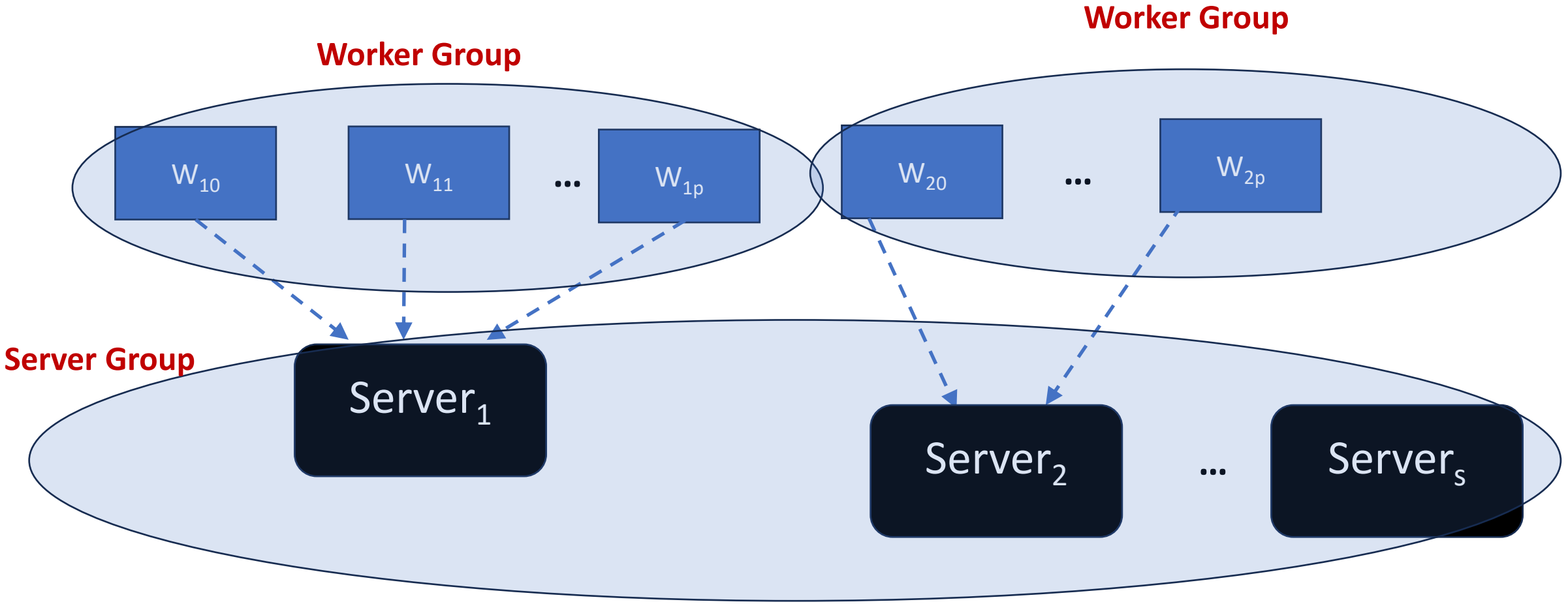
Distributed Systems and Failures

- Individual Nodes may fail frequently in distributed systems:
  - This is particularly so in commodity clusters
- Rate of node failure increases with the size of the system (*i.e., more nodes and more processes*)
  - **Cloud data centers** are made out of commodity clusters!
- Distributed Systems have to function (i.e. be available) despite node failures
  - This is referred to as fault tolerance and is achieved via
    - redundancy and failure recovery

Scalable and Dependable (*reliable and available*) architecture for ML

Each Worker Group includes a task scheduler.

Server Group must be fault tolerant!



Scalable and Dependable (*reliable and available*) architecture for ML

- This architecture is referred to as the ***parameter server*** model:
  - Different servers may handle different problems
  - Multiple servers may work on the same problem to improve performance
    - This will require additional combination/minimization processes and servers.
  - Multiple servers may work on the same problem for redundancy.

Parameter Server Model

- This model was popular for a few years
  - Google built an internal platform named DistBelief based on this model
    - DistBelief was optimized primarily for large clusters of multi-core nodes
    - GPU clusters were enabled later
  - Later, Google's TensorFlow provided programming flexibilities not available in DistBelief:
    - Adding new layers
    - Adding new ML training workflows
    - Optimizing or tuning ML algorithms

TensorFlow

- GPU acceleration has become a common tool for ML algorithms.
  - Building and testing on GPU workstations before scaling it to a GPU cluster is a common scenario as well.
- TensorFlow provides a unified programming interface and a common runtime on all these hardware platforms
  - while also supporting heterogeneous accelerators.
- e.g. Google's TPUs are special purpose accelerators for ML
  - that enable increased performance-per-watt compared to other state-of-the-art hardware.
- TensorFlow supports a common device abstraction for heterogeneous accelerators.

Gradient Descent

- One approach to error minimization is known as Gradient Descent:
  - Use the slope (i.e., gradient) of the loss function to update the parameters.
- This is particularly useful in Neural Networks in the back-propagation phase

Gradient Descent

- The gradient descent approach to minimize the error requires the following update to the parameters
  - $w = w - \eta * g(L, D, w)$
    - where g is the gradient function, L the loss function, and D, the dataset.
    - η denotes the learning rate – controls the amount by which the parameters are updated.
- The updates are done iteratively.

Gradient Descent

for i = 1 to num_iter :

1. grad = eval_gradient (loss_function , D , w)
2. w = w - learning_rate * grad

- This is Batch GD!
  - i.e., update is done after all the points in the dataset are considered.
- Batch Gradient Descent is slow to converge, particularly if same data (or similar) data is repeated within the dataset.:

Stochastic Gradient Descent (SGD)

1. for i = 1 to num_iter :
    1.1 shuffle (data)
    1.2 for example in data :
        1.2.1 grad = eval_gradient (loss_function , example , w)
        1.2.2. w = w - learning_rate * grad

This may converge faster but is completely sequential i.e., not easy to parallelize!

Mini-batch SGD

for i = 1 to num_iter:

1. shuffle (data)
2. for batch in get_batches (data , batch_size):
    1. grad = eval_gradient (loss_function , batch , w)
    2. w = w - learning_rate * grad

- Batches in the inner loop can be executed independently (locally)!
  - If necessary, batches can be obtained and stored locally at the start.
- Update has to be done in the parameter server

BITS Pilani

WILP

AIML CLZG516

ML System Optimization

Shan Sundar Balasubramaniam

innovate achieve lead

BITS Pilani

WILP

AIML CLZG516
ML System Optimization

Session 9: 20 Aug. 2023

System / Program Optimization Techniques:

- Locality-aware Programs, Massive Multi-threading
- Introduction to GPGPUs

Example: Matrix Multiplication

```
void matMult_IJK (float *a, float *b, float *c, int n)
{   // n x n row-major matrices a and b.
      for (int i = 0; i < n; i++)
      for (int j = 0; j < n; j++)
      {
            float temp = 0;
            for (int k = 0; k < n; k++)
                  temp += a[i*n+k]*b[k*n+j];
            c[i*n+j] = temp;
      }

}
```

```
void matMult_IKJ (float *a, float *b, float *c, int n)
{    // Sequential cache-aware algorithm
      // n x n row-major matrices a and b
      for (int i = 0; i < n; i++)
      {
         for (int j = 0; j < n; j++) { c[i*n+j] = 0;  }
         for (int k = 0; k < n; k++)
         for (int j = 0; j < n; j++)
            c[i*n+j] += a[i*n+k]*b[k*n+j];
      }

}
```

matMult_IKJ is faster than matMult_IJK (for large **n**): caching effect! (*see previous Session*)

- But matMult_IKJ accesses matrix **c** $O(N^3)$ times compared to **$O(N^2)$** accesses by matMult_IJK
- Both access **a** and **b** **$O(N^2)$** times

Implication: matMult_IKJ is better suited for large scale multi-threading!

Multi-threaded Matrix Multiplication

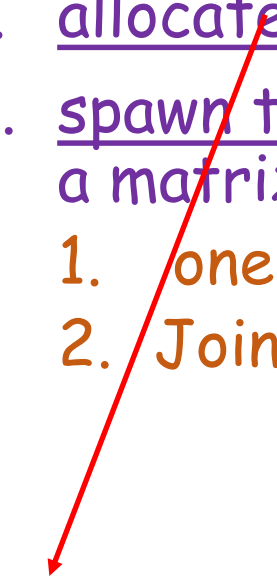
```
void MM_kernel (float *a, float *b, float *c, int n)
{       // thread code to compute one element of c
        // element (i,j) of the product to be computed
    int i = computeRowID();  // to be defined
    int j = computeColID(); // to be defined
    float temp = 0;
    for (int k = 0; k < n; k++)
            temp += a[i*n+k]*b[k*n+j];
    c[i*n+j] = temp;
}
```

1. allocate space for c
2. spawn threads in a grid (i.e., in a matrix)
    1. one thread per element of **c**
    2. Join the threads

Multi-threaded Matrix Multiplication

```
void MM_kernel (float *a, float *b, float *c, int n)
{       // thread code to compute one element of c
        // element (i,j) of the product to be computed
    int i = computeRowID();  // to be defined
    int j = computeColID(); // to be defined
    float temp = 0;
    for (int k = 0; k < n; k++)
            temp += a[i*n+k]*b[k*n+j];
    c[i*n+j] = temp;
}
```

1. allocate space for c
2. spawn threads in a grid (i.e., in a matrix)
    1. one thread per element of **c**
    2. Join the threads



space in shared memory

Multi-threaded Matrix Multiplication

```
void MM_kernel (float *a, float *b, float *c, int n)
{       // thread code to compute one element of c
        // element (i,j) of the product to be computed
    int i = computeRowID();  // to be defined
    int j = computeColID(); // to be defined
    float temp = 0;
    for (int k = 0; k < n; k++)
            temp += a[i*n+k]*b[k*n+j];
    c[i*n+j] = temp;
}
```

1. allocate space for c
2. spawn threads in a grid (i.e., in a matrix)
    1. one thread per element of **c**
    2. join the threads

O(n*n) threads:

- each thread computes c[i,j] for one specific (i,j)

- i and j computed from threadID

Multi-threaded Matrix Multiplication

```
void MM_kernel (float *a, float *b, float *c, int n)
{       // thread code to compute one element of c
        // element (i,j) of the product to be computed
    int i = computeRowID();  // to be defined
    int j = computeColID(); // to be defined
    float temp = 0;
    for (int k = 0; k < n; k++)
            temp += a[i*n+k]*b[k*n+j];
    c[i*n+j] = temp;
}
```

1. allocate space for c
2. spawn threads in a grid (i.e., in a matrix)
    1. one thread per element
    2. join the threads

Typically, computation may have to wait for all threads to finish before proceeding further

- in this example, it is a data-parallel computation
- *a.k.a* Single Program Multiple Data (SPMD) computation
- *a.k.a* Single Instruction Multi-Threading (SIMT)

Multi-threading in CPUs

- Multiple threads run in parallel by sharing resources:
  - In a single processor system:
    - Thread execution is interleaved – at a time one thread gets the processor
    - Threads share virtual memory:
      - Every thread gets its own (call) stack
      - but global data is shared via global memory or heap
    - Threads share physical memory
      - RAM
      - Caches
- Implication:
  - typically not more than a few threads can run in parallel without impacting performance!

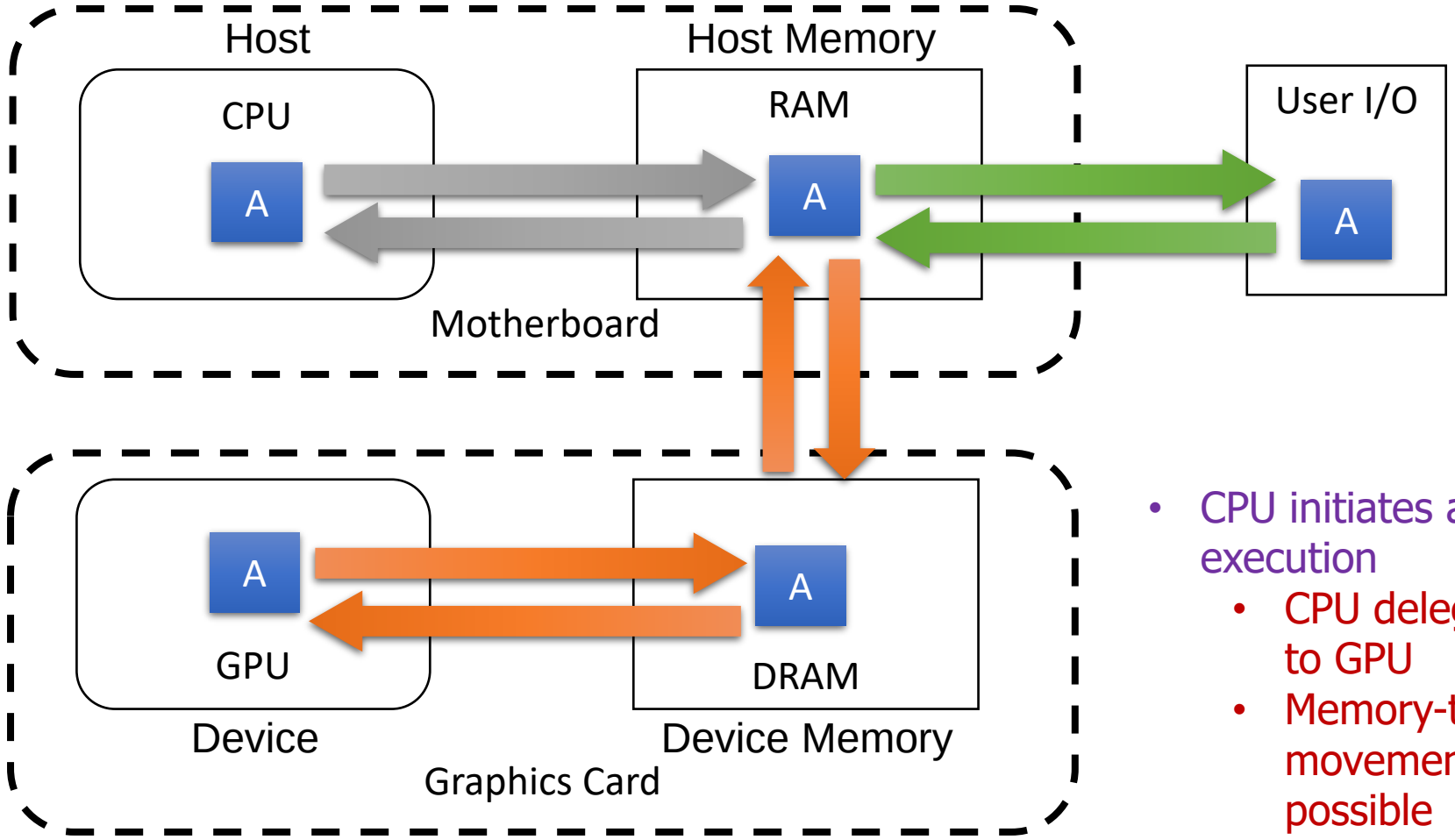
Multi-threading in CPUs: Multi-core processors

- Multiple threads run in parallel:
  - Each thread is scheduled to a core (at a time)
  - Cores are shared dynamically:
    - If there are **m** threads and **n** cores:
      - n-m cores are idle when n>m
      - More than one thread may be assigned to a core
        - i.e. threads interleave (like in the default scenario of one processor)
    - Threads share virtual memory:
    - Threads share physical memory:
      - RAM is shared
      - Cache is *private to a core* or *shared among the cores*

Introduction to GPGPUs

General Purpose Graphics Processing Units

GPU as a Co-processor



- CPU initiates and controls execution
  - CPU delegates work to GPU
  - Memory-to –memory movement of data possible

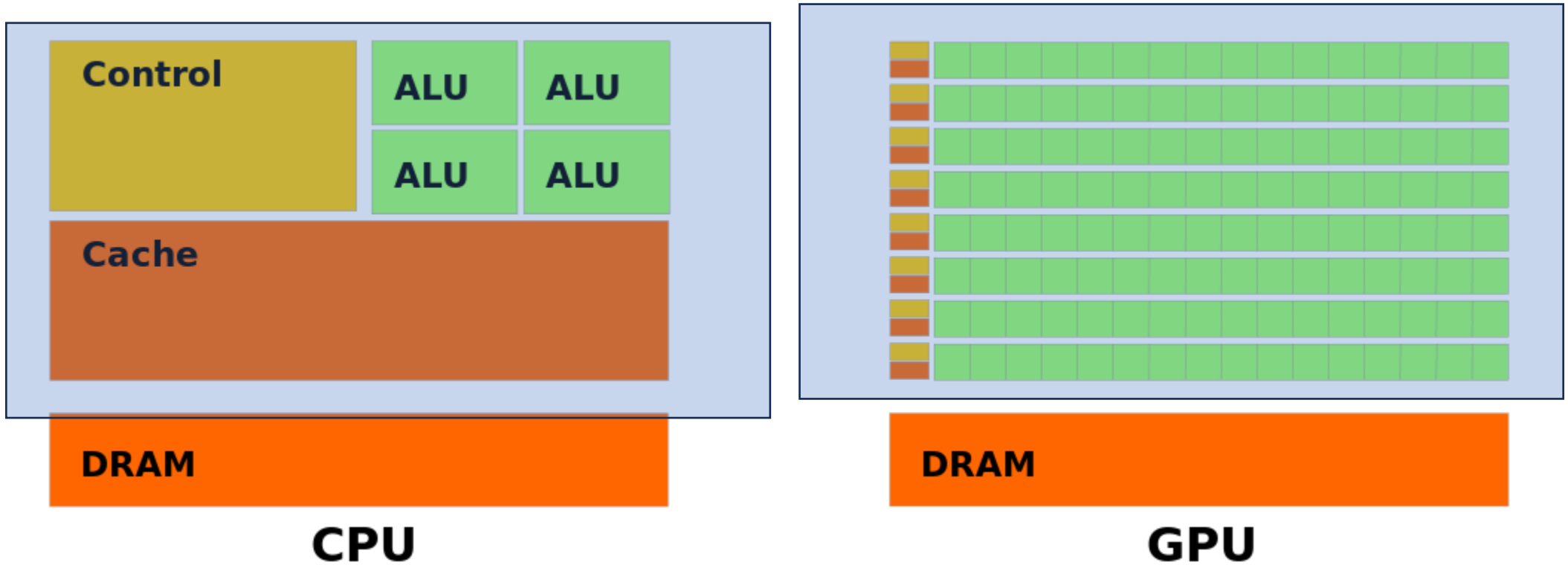
GPU Characteristics

- GPUs were designed for graphics processing where vertices and pixels are processed independently
  - Each processing step is usually arithmetic intensive
    - i.e., multiple operations are applied in between memory access
  - So, less control logic and more of arithmetic logic is required

- GPUs are designed for tasks that can tolerate high latency
  - as long as it can process a lot of tasks in one go
    - (i.e., high latency, high throughput)
- So, data caching is not a priority!

- More of the chip area can be given to ALUs
  - instead of Control Logic and Caches.
- Therefore, a lot of GPU threads (10s of thousands) can (should) execute at a time.

CPU Vs. GPU: Transistor Allocation Ratio



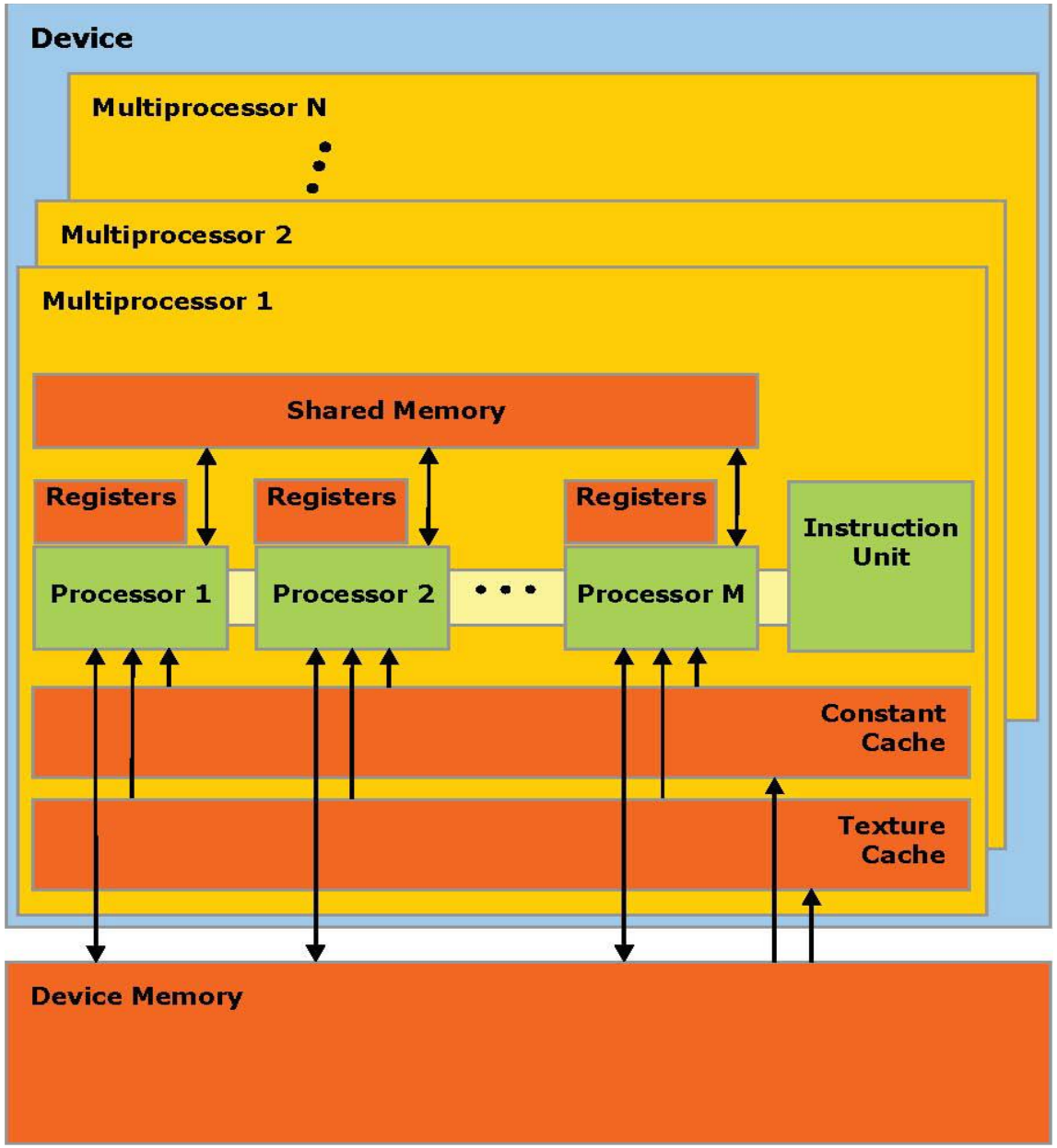
cf: **NVIDIA CUDA Programming Guide**

GPU Programming

- Kernel Programming Languages
  - Explicit and fine level control over threads and memory operations
  - Example:
    - CUDA (Compute Unified Driver Architecture) for Nvidia GPUs
      HIP for AMD and Nvidia GPUs
    - OpenCL for AMD and Nvidia GPUs
- Directive-based Programming Languages
  - No control over threads, compiler automatically generates code for parallelization
  - Example:
    - OpenMP for AMD and Nvidia GPUs
    - OpenACC for AMD and Nvidia GPUs

GPU Hardware

- Each GPU device has N streaming multiprocessors (SM)
  - each with M scalar processors (SP)
- e.g. NVIDIA Tesla C1060 (N=30, M=8)
  - 4GB off-chip device memory (global)
  - Each SM has 16KB of shared memory
  - Each SM has 16K registers (32-bit each)
- e.g. NVIDIA Tesla C2050 (N=14, M=32)
  - Each SM has 32K registers
  - Each SM has 64KB on-chip memory
    - divisible as 48KB + 16KB for shared memory and L1 cache (or the other way round)



GPU Multi-threading

- Single Instruction Multi-threading model
- Threads are light-weight:
  - Low context-switching overhead
- Thousands of threads can execute in parallel
  - Threads are grouped into blocks
    - 1D, 2D, or 3D (thread IDs)
      - Recall the matrix example.
  - Blocks are arranged in a grid
    - 1D, 2D, or 3D

Thread Execution

- When a kernel is launched:
  - blocks of the grid are enumerated and distributed to streaming multiprocessors (SM) with available capacity
- Each block is bound to one of the SM
  - Each block is divided into a group of 32 threads, known as a *Warp.*
  - Warp is the scheduling unit of .
- SM hardware implements zero-overhead Warp scheduling
  - Warps, whose next instruction has its operands ready for consumption,
    - are eligible for execution
  - Eligible Warps are selected for execution
    - on a prioritized scheduling policy
  - All threads in a Warp
    - execute the same instruction when selected

Thread Execution Support

- Instructions are pipelined
  - to leverage instruction-level parallelism
- Execution context for each warp
  - is maintained on-chip during the entire lifetime of the warp.
- Register file is partitioned among the warps.

(Back to) Matrix Multiplication

in GPGPUs

Multi-threaded Matrix Multiplication

GPU Kernel

```
void MM_kernel (float *a, float *b, float *c, int n)
{       // thread code to compute one element of c
        // element (i,j) of the product to be computed
    int i = computeRowID();
   int j = computeColID();
    float temp = 0;
    for (int k = 0; k < n; k++)
            temp += a[i*n+k]*b[k*n+j];
    c[i*n+j] = temp;
}
```

CPU code

1. allocate space for c
2. invoke threads in a grid (i.e., in a matrix)
    1. one thread per element of **c**

O(n*n) threads:

- each thread computes c[i,j] for one specific (i,j)

- i and j computed from blockID, threadID

Mapping and Grouping

- Suppose we tile an n * n matrix using tiles of dimensions p * q:
  - Assume p|n and q|n for convenience
- Tiles are indexed by
  - (u,v), 0 <= u < n/q, 0 <= v < n/p
- Example: Tiling a 16 * 16 matrix with 2 * 4 tiles

u

v

(0,0)	(1,0)	(2,0)	(3,0)
(0,1)	(1,1)	(2,1)	(3,1)
(0,2)			
(0,3)			
(0,4)			
(0,5)			
(0,6)			
(0,7)			

- *A thread may determine the co-ordinates of the block that it is part of using the variables* blockIdx.x and blockIdx.y
  i.e., (u,v) = (blockIdx.x, blockIdx.y)

- *The dimensions of a thread block may be obtained by using the variables* blockDim.x and blockDim.y
  i.e, p = blockDim.y and q = blockDim.x

- *The index of a thread within a block is*

(threadIdx.x, threadIdx.y)
where 0 <= threadIdx.x < q and 0 <= threadIdx.y < p

Multi-threaded Matrix Multiplication: Tiled version

GPU Kernel

```
void MM_kernel (float *a, float *b, float *c, int n)
{       // thread code to compute one element of c
        // element (i,j) of the product to be computed
 int i = blockIdx.y * blockDim.y + threadIdx.y;
int j = blockIdx.x * blockDim.x + threadIdx.x;
 float temp = 0;
    for (int k = 0; k < n; k++)
            temp += a[i*n+k]*b[k*n+j];
    c[i*n+j] = temp;
}
```

CPU code

```
allocate space for c
// ..
// define grid and block dimensions
dim3 grid (n/q, n/p);
dim3 block (q, p);
// invoke kernel
MM_kernel <<<grid, block>>> (d_A, d_B, d_C, n);
```

Tile/Block Dimensions

- How do you decide p and q?
  - Hardware limits the number of threads in a block
    - E.g. 512 in (some) Nvidia GPUs
  - Thus p*q <= 512

- Threads are scheduled in units of warps and a warp is of size 32.
  - Thus 32 | p*q
- This limits the options to
  - 8*8 or 16*16 if tiles have to be squares
  - 1*64, 1*128, 2*128, 4*128, 4*32, 8*32, 8*64, 16*32 … otherwise

Multi-threaded Matrix Multiplication

GPU Kernel

```
void MM_kernel (float *a, float *b, float *c, int n)
{       // thread code to compute one element of c
        // element (i,j) of the product to be computed
    int i = computeRowID();
   int j = computeColID();
    float temp = 0;
    for (int k = 0; k < n; k++)
            temp += a[i*n+k]*b[k*n+j];
    c[i*n+j] = temp;
}
```

CPU code

1. allocate space for c
2. invoke threads in a grid (i.e., in a matrix)
    1. one thread per element of **c**

Other options?

e.g. each thread computes a 1 * 4 sub-matrix of **c**

```
MatMult_kernel_1x4(float *a, float *b, float *c, int n)
{// thread to compute a 1 x 4 sub-matrix of c
// determine index of 1 x 4 c sub-matrix to compute
int i = blockIdx.y * blockDim.y + threadIdx.y;
int j = blockIdx.x * blockDim.x + threadIdx.x;
int nDiv4 = n/4;
int aNext = i*nDiv4;
int bNext = j;
float4 temp4;
temp4.x = temp4.y = temp4.z = temp4.w = 0;
```

```
for (int k = 0; k < nDiv4; k++)
{
    float4 aIn = a4[aNext++];  float4 bIn = b4[bNext];

    temp4.x += aIn.x*bIn.x; temp4.y += aIn.x*bIn.y;
    temp4.z += aIn.x*bIn.z; temp4.w += aIn.x*bIn.w;
    bNext += nDiv4;  bIn = b4[bNext];

    temp4.x += aIn.y*bIn.x; temp4.y += aIn.y*bIn.y;
    temp4.z += aIn.y*bIn.z; temp4.w += aIn.y*bIn.w;
    bNext += nDiv4;   bIn = b4[bNext];

    temp4.x += aIn.z*bIn.x; temp4.y += aIn.z*bIn.y;
    temp4.z += aIn.z*bIn.z; temp4.w += aIn.z*bIn.w;
    bNext += nDiv4;  bIn = b4[bNext];

    temp4.x += aIn.w*bIn.x; temp4.y += aIn.w*bIn.y;
    temp4.z += aIn.w*bIn.z; temp4.w += aIn.w*bIn.w;
    bNext += nDiv4;

  }
   c4[i*nDiv4+j] = temp4;
}
```

Memory Access Cost

- The costliest access is from device memory:
  - It is off-chip
    - and hence slower than on-chip memories
  - and data movement takes additional time
    - Due to bandwidth limitations of the bus and the pins
- Usually, a single access fetches multiple words
  - Typically 128 bytes (or four 32-byte words).
- So, unless the spatial locality of such accesses is very high,
  - bandwidth utilization will be low (i.e. fetch and throw).
- Alternative?
  - Use shared memory as much as possible.

innovate achieve lead

BITS Pilani

WILP

ML Systems Optimization

Shan Sundar Balasubramaniam

innovate achieve lead

BITS Pilani

WILP

ML Systems Optimization

• Shan Sundar Balasubramaniam

- *Matrix Multiplication in GPUs*
- *CNN*
  - *Gradient Descent*
  - *Convolution Layers and Fully Connected Layers*

(Back to) Matrix Multiplication

in GPGPUs

Multi-threaded Matrix Multiplication

GPU Kernel

```
void MM_kernel (float *a, float *b, float *c, int n)
{       // thread code to compute one element of c
        // element (i,j) of the product to be computed
    int i = computeRowID();
   int j = computeColID();
    float temp = 0;
    for (int k = 0; k < n; k++)
            temp += a[i*n+k]*b[k*n+j];
    c[i*n+j] = temp;
}
```

CPU code

1. allocate space for c
2. invoke threads in a grid (i.e., in a matrix)
    1. one thread per element of **c**

O(n*n) threads:
- each thread computes c[i,j] for one specific (i,j)
- i and j computed from blockID, threadID

Alternatives?

Mapping and Grouping

- Suppose we tile an n * n matrix using tiles of dimensions p * q:
  - Assume p|n and q|n for convenience
- Tiles are indexed by
  - (u,v), 0 <= u < n/q, 0 <= v < n/p
- Example: Tiling a 16 * 16 matrix with 2*4-tiles

u

v

(0,0)	(1,0)	(2,0)	(3,0)
(0,1)	(1,1)	(2,1)	(3,1)
(0,2)	(1,2)	(2,2)	(3,2)
(0,3)	(1,3)	(2,3)	(3,3)
(0,4)	(1,4)	(2,4)	(3,4)
(0,5)	(1,5)	(2,5)	(3,5)
(0,6)	(1,6)	(2,6)	(3,6)
(0,7)	(1,7)	(2,7)	(3,7)

- *A thread may determine the co-ordinates of the block that it is part of using the variables* blockIdx.x and blockIdx.y
  i.e., (u,v) = (blockIdx.x, blockIdx.y)
- *The dimensions of a thread block may be obtained by using the variables* blockDim.x and blockDim.y
  i.e, p = blockDim.y and q = blockDim.x
- *The index of a thread within a block is*

(threadIdx.x, threadIdx.y)
where 0 <= threadIdx.x < q and 0 <= threadIdx.y < p

Tile/Block Dimensions

- How do you decide p and q?
  - Hardware limits the number of threads in a block
    - E.g. 512 in (some) Nvidia GPUs
  - Thus p*q <= 512

- Threads are scheduled in units of warps and a warp is of size 32.
  - Thus 32 | p*q
- This limits the options to
  - 8*8 or 16*16 if tiles have to be squares
  - 1*64, 1*128, 2*128, 4*128, 4*32, 8*32, 8*64, 16*32 … otherwise

Multi-threaded Matrix Multiplication: Tiled version

GPU Kernel

```
void MM_kernel (float *a, float *b, float *c, int n)
{       // thread code to compute one element of c
        // element (i,j) of the product to be computed
 int i = blockIdx.y * blockDim.y + threadIdx.y;
int j = blockIdx.x * blockDim.x + threadIdx.x;
 float temp = 0;
    for (int k = 0; k < n; k++)
            temp += a[i*n+k]*b[k*n+j];
    c[i*n+j] = temp;
}
```

CPU code

```
allocate space for c
// ..
// define grid and block dimensions
dim3 grid (n/q, n/p);
dim3 block (q, p);
// invoke kernel
MM_kernel <<<grid, block>>> (d_A, d_B, d_C, n);
```

Multi-threaded Matrix Multiplication

GPU Kernel

```
void MM_kernel (float *a, float *b, float *c, int n)
{       // thread code to compute one element of c
        // element (i,j) of the product to be
computed
    int i = computeRowID();
   int j = computeColID();
    float temp = 0;
    for (int k = 0; k < n; k++)
            temp += a[i*n+k]*b[k*n+j];
    c[i*n+j] = temp;
}
```

CPU code

1. allocate space for c
2. invoke threads in a grid (i.e., in a matrix)

   each thread computes a **1*4** sub-matrix of **c**

```
MatMult_kernel_1x4(float *a, float *b, float *c, int n)
{// thread to compute a 1 x 4 sub-matrix of c
// determine index of 1 x 4 c sub-matrix to compute
int i = blockIdx.y * blockDim.y + threadIdx.y;
int j = blockIdx.x * blockDim.x + threadIdx.x;
int nDiv4 = n/4;
int aNext = i*nDiv4;
int bNext = j;
float4 temp4;
temp4.x = temp4.y = temp4.z = temp4.w = 0;
```

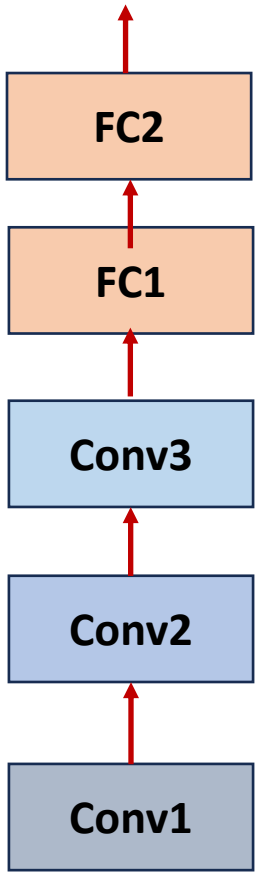
```
for (int k = 0; k < nDiv4; k++)
{
    float4 aIn = a4[aNext++];  float4 bIn = b4[bNext];

    temp4.x += aIn.x*bIn.x; temp4.y += aIn.x*bIn.y;
    temp4.z += aIn.x*bIn.z; temp4.w += aIn.x*bIn.w;
    bNext += nDiv4;  bIn = b4[bNext];

    temp4.x += aIn.y*bIn.x; temp4.y += aIn.y*bIn.y;
    temp4.z += aIn.y*bIn.z; temp4.w += aIn.y*bIn.w;
    bNext += nDiv4;   bIn = b4[bNext];

    temp4.x += aIn.z*bIn.x; temp4.y += aIn.z*bIn.y;
    temp4.z += aIn.z*bIn.z; temp4.w += aIn.z*bIn.w;
    bNext += nDiv4;  bIn = b4[bNext];

    temp4.x += aIn.w*bIn.x; temp4.y += aIn.w*bIn.y;
    temp4.z += aIn.w*bIn.z; temp4.w += aIn.w*bIn.w;
    bNext += nDiv4;

  }
   c4[i*nDiv4+j] = temp4;
}
```

Memory Access Cost

- The costliest access is from device memory:
  - It is off-chip
    - and hence slower than on-chip memories
  - and data movement takes additional time
    - Due to bandwidth limitations of the bus and the pins
- Usually, a single access fetches multiple words
  - Typically 128 bytes (or four 32-byte words).
- So, unless the spatial locality of such accesses is very high,
  - bandwidth utilization will be low (i.e. fetch and throw).
- Alternative?
  - Use shared memory as much as possible.

CNN – Implementation and (Code) Optimization

CNN Layers - Example

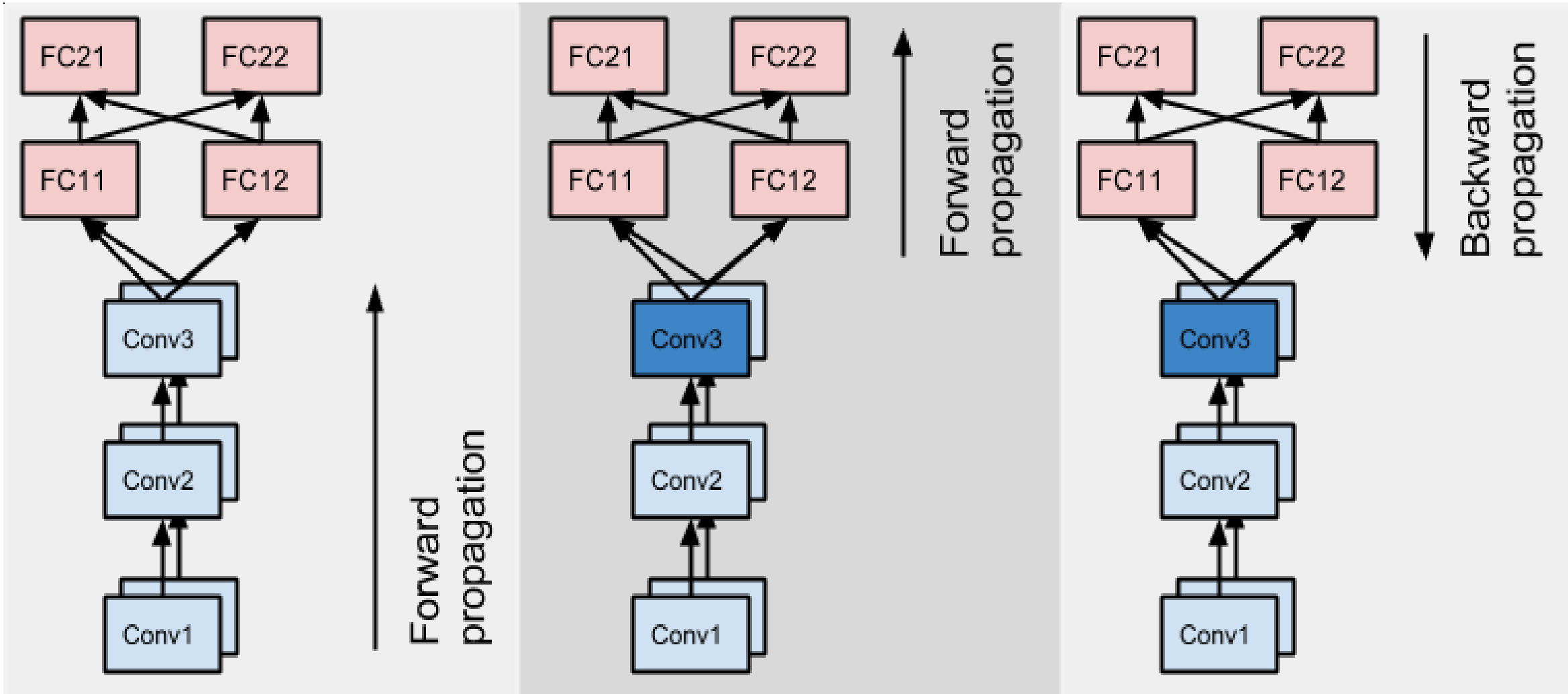


3 Convolution layers and 2 Fully Connected layers

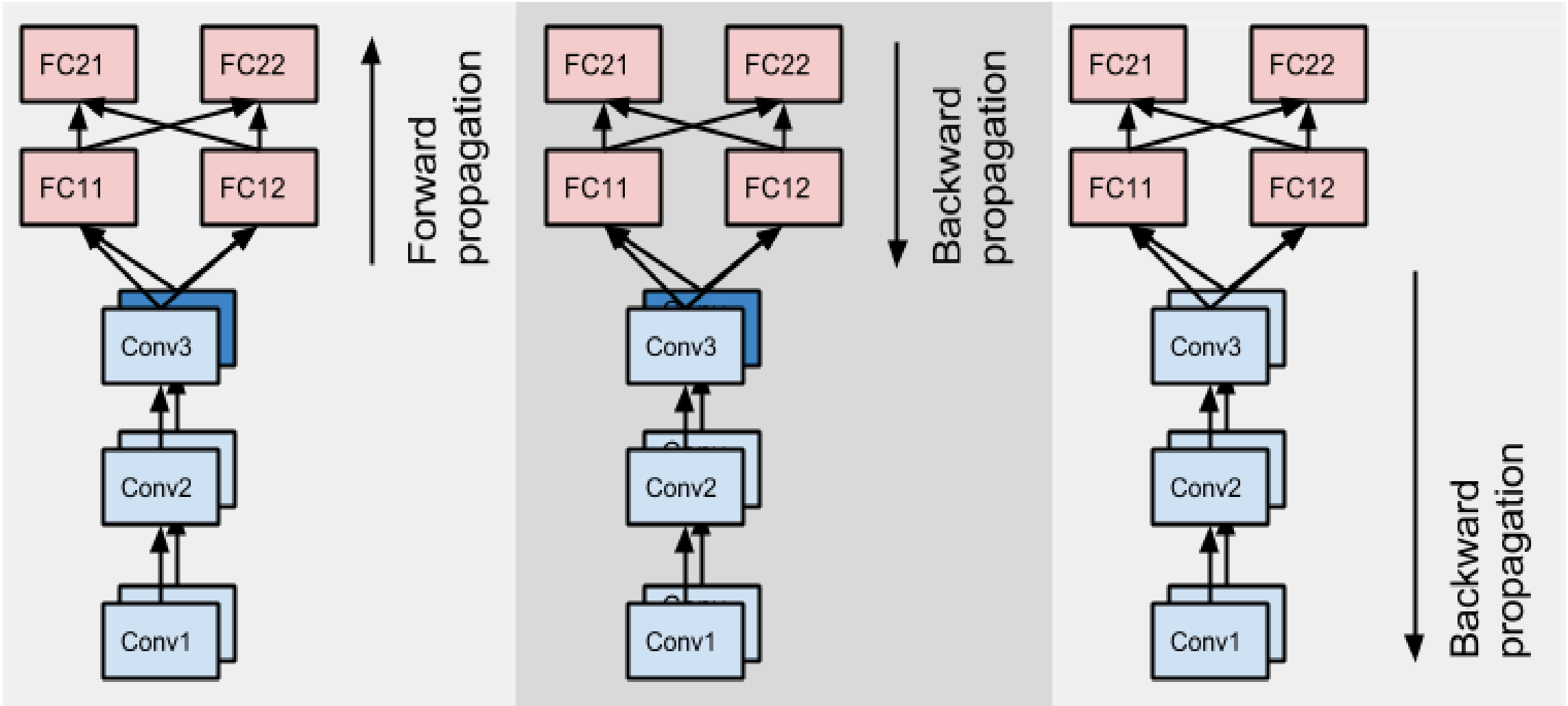
CNN computations

- Gradient Descent
  - – we will deal with this separately.
  - Not usually done in GPGPUs
- Fully Connected Layers
  - ~ 5 to 10% of the computations
  - Model parallelism
    - All workers (threads) train on the same batch
- Convolutional Layers
  - ~ 5 to 10% of the computations
  - Each worker (thread) trains on the same convolutional layers on different (data) batches

CNN Layers - Parallel (2 workers)



CNN Layers - Parallel (2 workers)



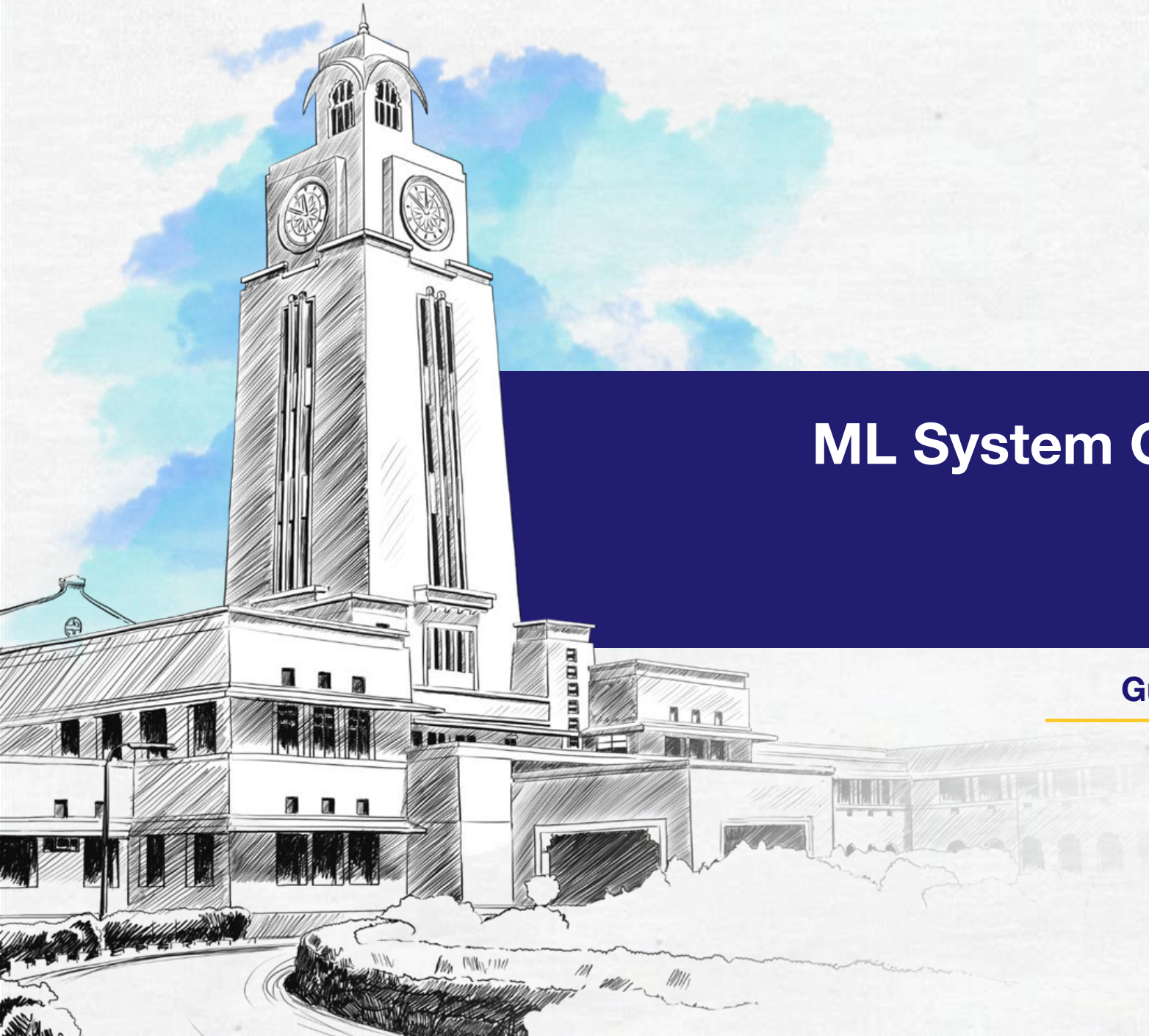
Convolution Kernel

$$\text{Out}[n, k, y, x] = \Sigma_{c,fx,fy}\ \text{In}[n, c, y+f\ y, x+f\ x] \times \text{Ker}\ [k, c, f\ y, f\ x]$$

CNN Kernel – (2D Convolution) - Untiled code

```
for k in range(0, K):
    for x in range(0, X):
        for y in range(0, Y):
            for c in range(0, C):
                for fy in range(0, Fy):
                    for fx in range(0, Fx):
                        Out[k,y,x]+=In[c,y+fy,x+fx]
                                    x Ker[k,c,fy,fx]
```

```
/ Grid level tiling
orall bc in range(0, C, B^C):
  forall bk in range(0, K, W^k):
    forall bx in range(0, X, W^x):
      forall by in range(0, Y, W^y):
        for c in range(bc, bc + B^c, W^c):
        // shared memory loading
        I_sm = In [c, by: by + W^y + F_y−1, bx: bx + W^x + F_x−1]
        K_sm = Ker[bk: bk + W^k, c, F_y, F_x]
        // Thread block level tiling
        forall wk in range(bk, bk + W^k, T^k):
          forall wx in range(bx, bx + W^x, T^x):
            forall wy in range(by, by + W^y, T^y):
            // Warp level tiling
            forall tk in range(wk, wk + T^k, R^k):
              forall tx in range(wx, wx + T^x, R^x):
                forall ty in range(wy, wy + T^y, R^y):
                // Register Tiling
```



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

ML System Optimization - Session 11

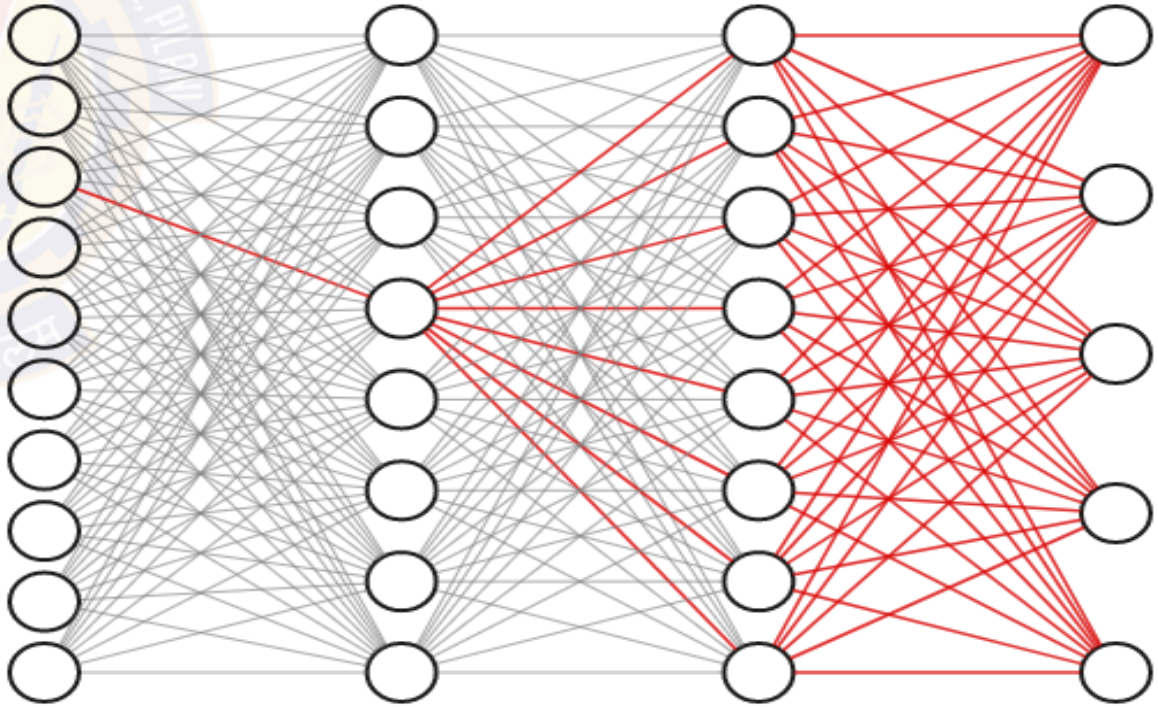
Model Compression

Guest Lecture by Prof. Anita Ramachandran

BITS, Pilani

Agenda

- **Why model compression**
- Model compression techniques
- Measuring goodness of model compression
- Which technique to use
- Challenges & Trade-offs



Measuring the performance of a deployed ML inference system

- Accuracy
- Latency
- Throughput
- Energy
- Model size
- Memory consumption
- Cost

Trade-offs

Why Model Compression

- Larger NN models – better performance, however:
  - More storage space – harder to distribute
  - Higher latency for each forward pass in training + inference time
  - May need more expensive hardware
  - Limited usage in applications where test sets are large (e.g. Google), where storage space is at a premium (e.g. PDAs), or where computational power is limited (e.g. hearing aids)
- Bigger models complicates democratization of AI models to resource constrained environments

- ML for edge
  - Industry 4.0, surveillance, intelligent transport system, health, energy management
  - Challenges
  - Techniques for edge intelligence
    - Federated learning
    - Model partitioning
    - Edge pre-processing
    - Scheduling
    - Cloud pre-training
    - Edge-only
    - Model compression

Agenda

- Why model compression
- **Model compression techniques**
- Measuring goodness of model compression
- Which technique to use
- Challenges

Model Compression Techniques

- Old school compression (Huffman coding or gzip)
- Parameter pruning and quantization
  - Quantization
  - Network pruning
- Low-rank factorization
- Knowledge distillation

Model compression techniques are lossy

Quantization

- Network quantization compresses the original network by reducing the number of bits required to represent each weight
- Have shown to significantly reduce memory usage and float point operations with little loss in classification accuracy
- Binary weight neural network
  - 1 bit representation of each weight
  - E.g. BinaryConnect, BinaryNet, XNOR
  - Drawbacks:
    - Lower accuracy when dealing with large CNNs such as GoogleNet
    - Binarization techniques are based on simple matrix approximations – effect on accuracy loss is ignored

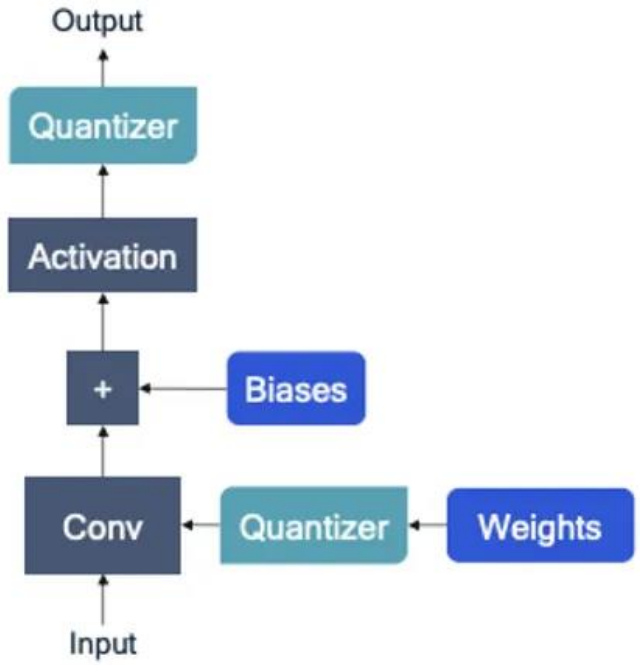
Data type	Represented Value	% Deviation	Storage bits
FP64	3.141592653589793	-	64 bits
FP32	3.141592653	5.97e-09 %	32 bits
FP16	3.1415	9.39e-04 %	16 bits
INT8	3	4.5 %	8 bits

Pros

- Can fit more numbers in memory
- Can store more numbers in cache
- Can transmit more numbers per second
- Less energy

Quantization

- What to quantize
  - Weights, Activations, Loss computation, backprop
- When to quantize
  - Post Training Quantization
  - Quantization Aware Training
- How to quantize
- How much to quantize

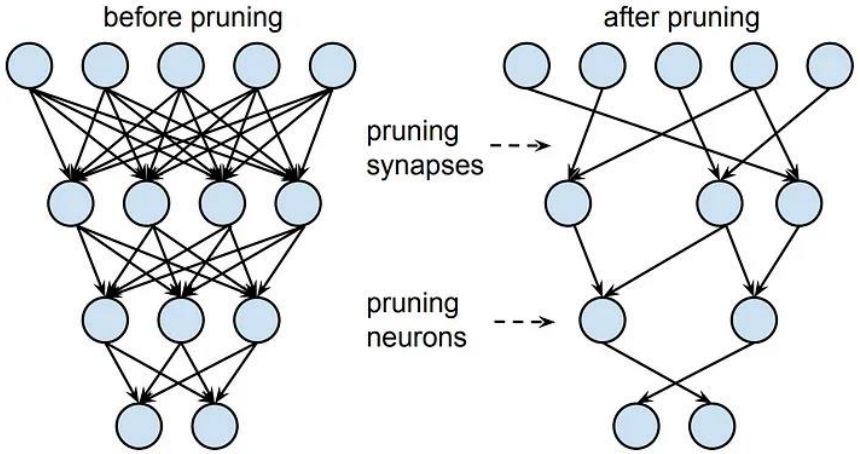


Model	Top-1 Accuracy (Original)	Top-1 Accuracy (Post Training Quantized)	Top-1 Accuracy (Quantization Aware Training)	Latency (Original) (ms)	Latency (Post Training Quantized) (ms)	Latency (Quantization Aware Training) (ms)	Size (Original) (MB)	Size (Optimized) (MB)
Mobilenet-v1-1-224	0.709	0.657	0.70	124	112	64	16.9	4.3
Mobilenet-v2-1-224	0.719	0.637	0.709	89	98	54	14	3.6
Inception_v3	0.78	0.772	0.775	1130	845	543	95.7	23.9
Resnet_v2_101	0.770	0.768	N/A	3973	2868	N/A	178.3	44.9

Table 1 Benefits of model quantization for select CNN models

Network Pruning

- NN Pruning - Remove synapses or neurons
  - Remove weights
    - Setting individual parameters to zero to make the network sparse
    - Remove all weights below a certain threshold
    - Reduces no: of parameters in the model; no change to architecture
  - Remove neurons
    - Remove neurons from the network – changes architecture
    - Based on activations on training data
    - Based on redundancy of parameters



Network Pruning

- What to prune
- When to prune
  - Static pruning, dynamic pruning
- How much to prune
  - Brute force, OBD
- How to prune
  - recoverable pruning
- Combine with fine-tuning
  - Fine-tuning: re-training a previously trained model
    - Retraining the network starting from some given parameters by running a relatively small number of epochs with a relatively low learning rate
  - Prune and then fine-tune
  - Fine-tuning operates on those weights that were not pruned
  - Can be done for low precision models also
- Caveats:
  - Pruning with L1 or L2 regularization requires more iterations to converge than general methods
  - All pruning criteria require manual setup of sensitivity for layers, which demands fine-tuning of the parameters and could be cumbersome for some applications
  - Network pruning usually is able to reduce model size but not improve the efficiency (training or inference time)

Demo – Weight Pruning

- Train a tf.keras model for MNIST from scratch.
  - Baseline test accuracy: 0.978600025177002
  - Saved baseline model to: /tmpfs/tmp/tmpzufz19sr.h5
- Fine tune the model by applying the pruning API and see the accuracy.
  - UpdatePruningStep during training
  - Total params: 40,805; Trainable params: 20,410; Non-trainable params: 20,395
  - Baseline test accuracy: 0.978600025177002
  - Pruned test accuracy: 0.9702000021934509
- Create 3x smaller TF and TFLite models from pruning.
  - Strip-pruning
  - TFLiteConverter
  - Size of gzipped baseline Keras model: 78186.00 bytes
  - Size of gzipped pruned Keras model: 25841.00 bytes
  - Size of gzipped pruned TFlite model: 24940.00 bytes
- Create a 10x smaller TFLite model from combining pruning and post-training quantization.
  - By applying post-training quantization to the pruned model (tf.lite.Optimize.DEFAULT)
- See the persistence of accuracy from TF to TFLite.
  - Pruned and quantized TFLite test_accuracy: 0.9699
  - Pruned TF test accuracy: 0.9702000021934509

Low Rank Factorization

- Convolution operations contribute to the bulk of most computations -> reduce the convolution layer to improve the compression rate in deep DNNs
- $A=BC^T$ → Rank factorization; helps to compactly store A by storing its factors B and C
- Where A – mxn, B – mxr, C^T – rxn
- No: of elements in A = mn
- No: of elements in BC^T = mr + nr (< mn)
- Eg: If A is a 50x100 matrix with rank(A) = 20, no: of elements in A = 5000 and no: of elements in BC^T = 1000 + 2000 = 3000
- Lower the rank, more the compression
- Computing rank factorization
  - Using any matrix factorization – LU, QR, Eigenvalue decomposition, SVD, several variants
- Role of approximation

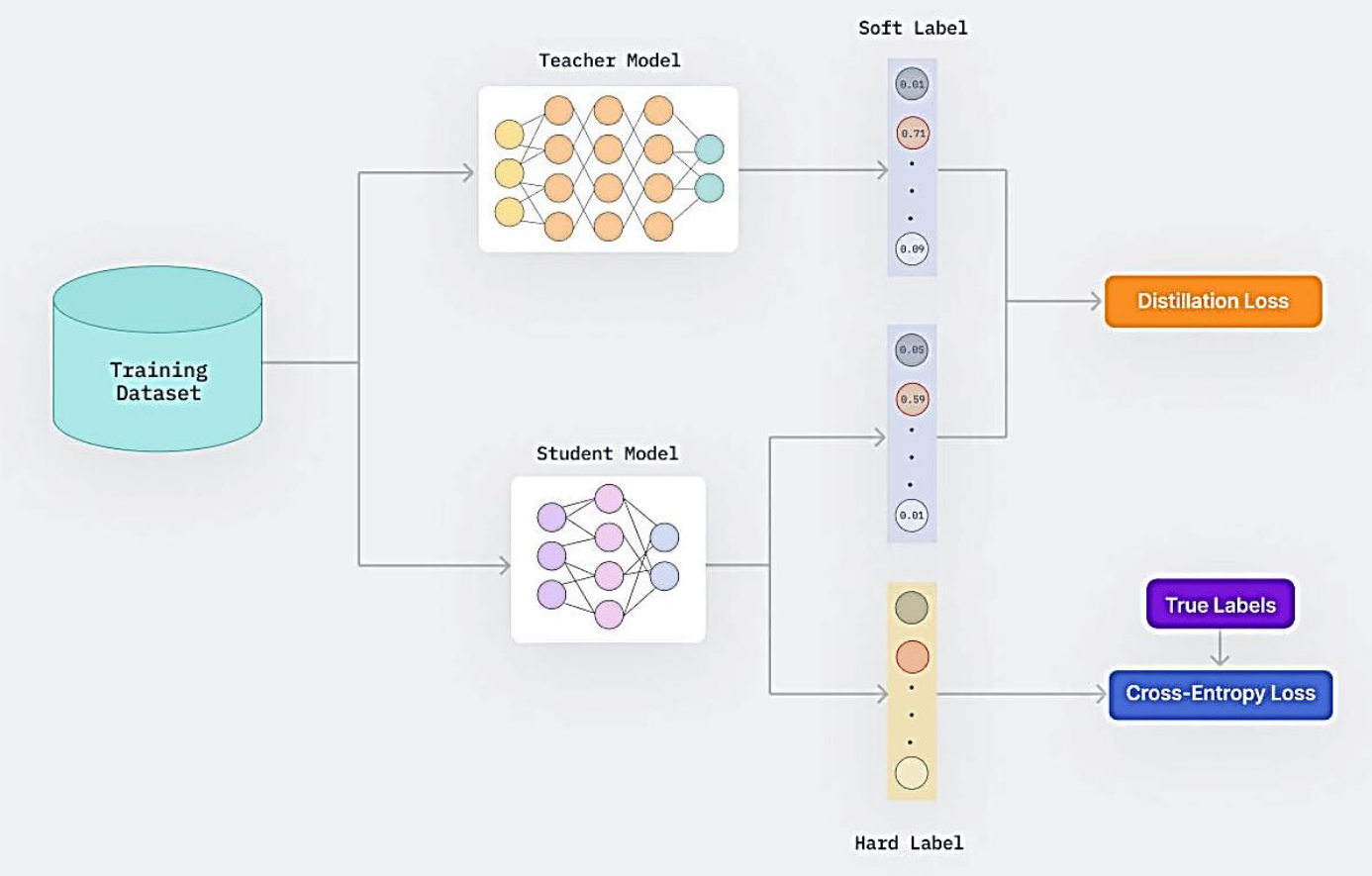
Low Rank Factorization

- Many matrices in applications are not low rank
- Solution - approximate the high-rank matrix with a low-rank one
  - Eg: r-truncated singular value decomposition
- Caveats
  - Not every matrix is well-approximated by one of small rank
  - Approximation errors and its propagations through further steps of the algorithm need to be analyzed and controlled along with other sources of error in the procedure
  - Decomposition operations are computationally expensive
  - Current methods perform low-rank approximation layer by layer, thus cannot perform global parameters compression, which is important as different layers hold different information
  - Requires extensive model retraining to achieve convergence when compared to the original model

Knowledge Distillation

- Shift knowledge from a large teacher model into a small one by learning the class distributions output via softmax
- Compress deep and wide networks into shallower ones, where the compressed model mimics the function learned by the complex model
- Distillation of knowledge - knowledge is transferred from the teacher network to the student network through a loss function where the optimization target is to match the class-wise probability distribution of the student network to the probability output by the teacher
  - This concept of model compression was generalized, and the concept of distillation was formulated in 2015 by Hinton et al., “Distilling the Knowledge in a Neural Network”.

Knowledge Distillation - Process

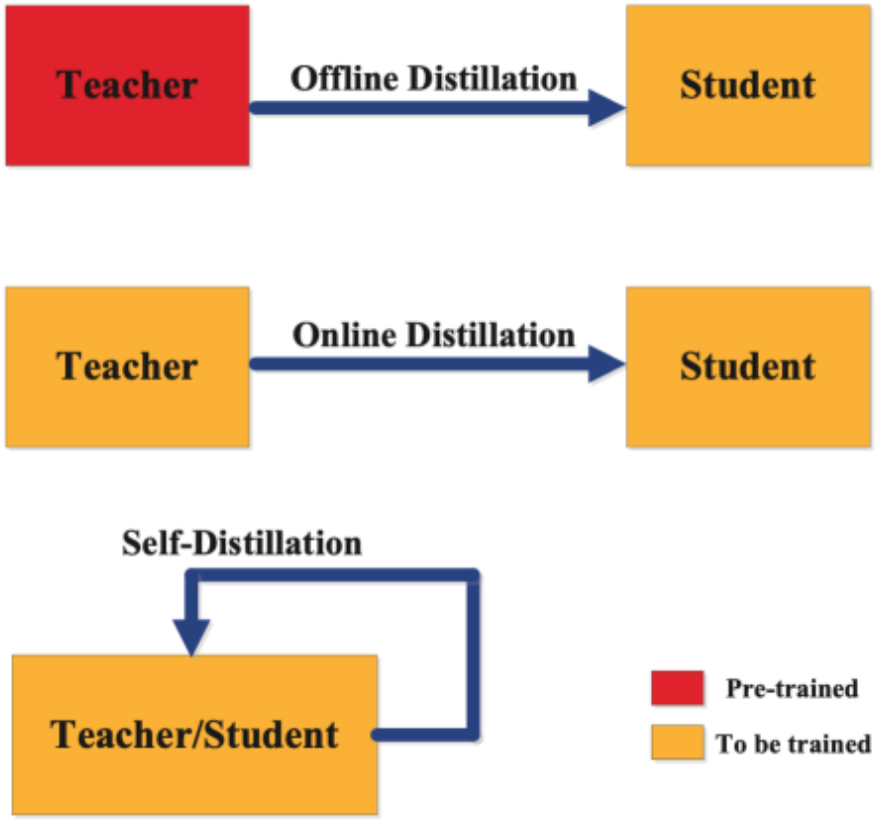


- How does it work?
  - Use the class prediction probabilities generated by the teacher network as “soft targets” for training the student model
  - If the teacher model is an ensemble, use the arithmetic or geometric mean of their individual predictive distributions as the soft targets

Knowledge Distillation - Principles

- Knowledge in NNs
- Response-based Knowledge systems
  - Knowledge - The information is obtained from the output layer of the teacher model
  - In such models, the student model is expected to mimic the logits (class probabilities) of the teacher model predictions
- Feature-based Knowledge systems
  - Knowledge - Feature maps from the intermediate layers of a network in the teacher model
  - The intermediate layers learn to discriminate specific features and this knowledge can be used to train a student model.
  - The goal is to train the student model to learn the same feature activations as the teacher model
  - The distillation loss function achieves this by minimizing the difference between the feature activations of the teacher and the student models.
- Relation-based Knowledge systems
  - Knowledge - The relationship between feature maps in the teacher model
  - This relationship can be modeled as correlation between feature maps, graphs, similarity matrix, feature embeddings, or probabilistic distributions based on feature representations.

Schemes for Knowledge Distillation



Knowledge Distillation – Programming Logic (Keras)

- Create the teacher and student models, prepare the datasets, train the teacher (specify optimizer, loss function, metrics), distil teacher to student
  - Optimizers provide us with accelerated gradient descent for unconstrained optimization problems
  - Several algorithms
  - The loss function for NNs is non-convex. With optimizers, affinity towards local minima is different. The gradient is reduced in the same direction as the previous reduction of gradient.
- Distiller class in Keras
  - Do forward pass of the teacher to get teacher predictions, forward pass of the student to get student predictions, compute student loss (ground truth vs student predictions) and distillation loss (teacher predictions vs student predictions), compute total loss, compute the gradients, update the weights using the optimizer, update the performance metrics

Knowledge Distillation - More

- Applications
  - Used for distilling ensemble models
  - Computer vision (image classification, face recognition, pose estimation, video captioning), NLP (neural machine translation, text generation), speech (speech synthesis, speaker recognition)

- Caveats:
  - KD can only be applied to tasks with softmax loss function, which hinders its usage.
  - KD-based approaches generally achieve less competitive performance compared with other type of approaches

Model Compression Techniques

- Parameter pruning and quantization
  - Quantization and binarization
  - Network pruning
- Low-rank factorization
- Knowledge distillation

Theme Name	Description	Applications	More Details
Parameter pruning and sharing	Reducing redundant parameters that are not sensitive to the performance	Convolutional layer and fully connected layer	Robust to various settings, can achieve good performance, can support both training from scratch and pretrained model
Low-rank factorization	Using matrix/tensor decomposition to estimate the informative parameters	Convolutional layer and fully connected layer	Standardized pipeline, easily implemented, can support both training from scratch and pretrained model
Transferred/compact convolutional filters	Designing special structural convolutional filters to save parameters	Only for convolutional layer	Algorithms are dependent on applications, usually achieve good performance, only support training from scratch
KD	Training a compact neural network with distilled knowledge of a large model	Convolutional layer and fully connected layer	Model performances are sensitive to applications and network structure, only support training from scratch

Agenda

- Why model compression
- Model compression techniques
- **Measuring goodness of model compression**
- Which technique to use
- Challenges & Trade-offs

Measuring Quality of Model Compression & Acceleration

- Compression rate $\alpha(M, M^*) = \frac{a}{a^*}$.
- Index space saving $\beta(M, M^*) = \frac{a - a^*}{a^*}$.
- Speedup rate $\delta(M, M^*) = \frac{s}{s^*}$.
  - a is the number of the parameters in the original model M and a* is that of the compressed model M*, s is the running time of M and s* is that of M*
- In general, the compression rate and speedup rate are highly correlated, as smaller models often results in faster computation for both the training and the testing stages

Agenda

- Why model compression
- Model compression techniques
- Measuring goodness of model compression
- **Which technique to use**
- Challenges & Trade-offs

Which Technique to Use

- For different applications with different CNN designs, the relation between parameter size and computational time might be different
- Which technique to use - Some considerations
  - If the applications need compacted models from pretrained deep nets, you can choose either pruning & quantization or low rank factorization based methods
  - If you need end-to-end solutions for your problem, the low rank and transferred convolutional filters approaches should be considered
  - The approaches of pruning & quantization generally give reasonable compression rate while not hurting the accuracy. Thus for applications which requires stable model performance, it is better to utilize pruning & quantization
  - If your application involves small/medium size datasets or requires significantly improving efficiency, you can try the knowledge distillation approaches. The compressed student model can take the benefit of transferring knowledge from teacher model, achieving robust performance when datasets are not large.
- These techniques are orthogonal
  - It is reasonable to combine two or three of them to maximize the gain.
  - E.g., for object detection, which requires both convolutional and fully connected layers, you can compress the convolutional layers with a low rank based method and the fully connected layers with a pruning technique.

Which Technique to Use

- Hardware dependencies
  - CPUs, GPUs [Accelerators]
  - Model performance differs
    - E.g. if the accelerator uses low-precision arithmetic, the performance of an algorithm in full precision on a CPU may differ from the performance of the algorithm in low-precision on the accelerator
- How to choose the hardware?
  - Purpose – training or inference?
  - Budget?
  - Cloud or edge?
  - Throughput? Latency? Power consumption?

Agenda

- Why model compression
- Model compression techniques
- Measuring goodness of model compression
- Application specific nuances
- Which technique to use
- **Challenges & Trade-offs**

Challenges

- Most of the current state-of-the-art approaches build on well-designed CNN models, which have limited freedom to change the configuration (e.g., network architectures, hyper-parameters)
- Hardware constraints in various small platforms (e.g., mobile, robotic, self-driving car) are still a major problem to hinder the extension of deep CNNs.
- Methods of structural matrix and transferred convolutional filters impose prior human knowledge to the model, which could significantly affect the performance and stability. It is critical to investigate how to control the impact of those prior knowledge
- The black box mechanism is still the key barrier to the adoption. For example, why some neurons/connections are pruned is not clear. Exploring the knowledge interpretability is still an important challenge

Trade-Offs

- Model performance, latency, memory, power consumption (sensing, computation, communication), cost

- Eg. If we were to deploy ChatGPT on mobile phones

  - Option 1 - Deploy a compressed ChatGPT model on the phone
    - Local inferencing, less response time
    - Extra memory requirements on the phone

  - Option 2 - Query a server that hosts the ChatGPT model
    - Remote inferencing, more data exchange/network traffic, more response time
    - No additional memory required on the phone
    - Power consumption may increase because of the network communication overhead

  - There could be intermediate options too

Thank You!

Storing Sparse Vectors/Matrices

$$\begin{bmatrix} 3 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 7 & 0 \end{bmatrix}$$

indexes: $\begin{bmatrix} 0 & 4 & 8 \end{bmatrix}$

values: $\begin{bmatrix} 3 & 2 & 7 \end{bmatrix}$

$$\begin{bmatrix} 5 & 0 & 0 & 3 & 0 & 1 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 3 & 0 \end{bmatrix}$$

row indexes: $\begin{bmatrix} 0 & 0 & 0 & 1 & 2 & 2 & 2 \end{bmatrix}$

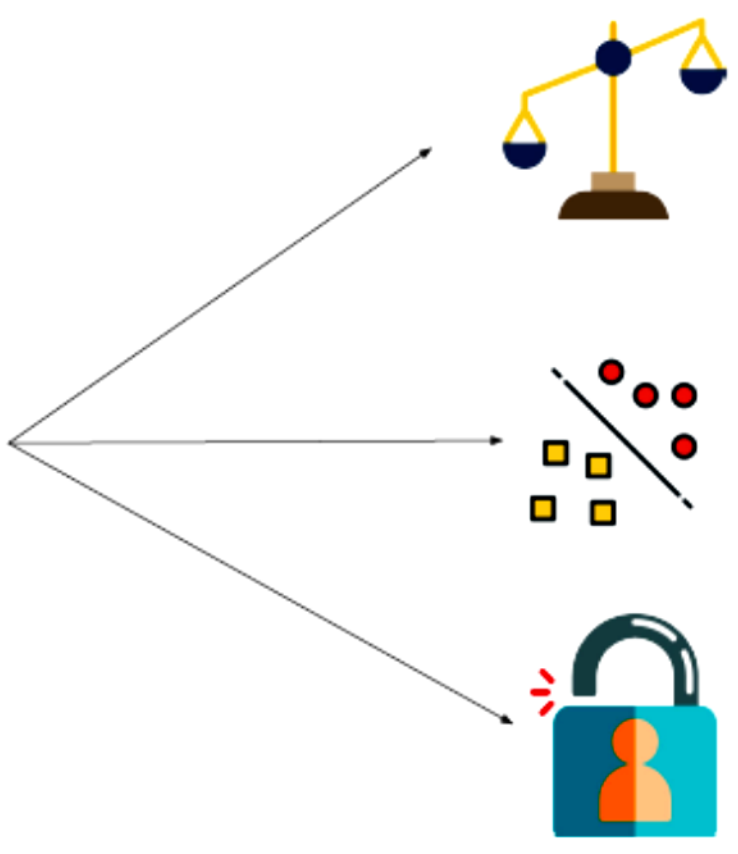
column indexes: $\begin{bmatrix} 0 & 3 & 5 & 1 & 0 & 1 & 4 \end{bmatrix}$

values: $\begin{bmatrix} 5 & 3 & 1 & 4 & 1 & 2 & 3 \end{bmatrix}$

More..

It is unrealistic to assume optimizing for one property holds all others static.

How we often talk about different properties in the literature.



Fairness - imposes constraint on optimization that reflects societal norms of what is fair.

Model Compression - compact machine learning models to work in resource constrained environments.

Model fragility and security - deploy secure models that protect user privacy.

ML System Optimization

Edge Intelligence – Case Studies

Anita Ramachandran

Why Edge

Overheads – bandwidth, latency, energy

Offload computation

More complexities

- Large DNN models
- Resource constrained devices

Cloud – long term and complex processing, storage

Edge – local algorithms, short term storage

Case Study 1 - Energy-Efficient Approximate Edge Inference Systems

Ref: Soumendu Kumar Ghosh, Arnab Raha, and Vijay Raghunathan. 2023. Energy-Efficient Approximate Edge Inference Systems. ACM Trans. Embed. Comput. Syst. 22, 4, Article 77 (July 2023), 50 pages. https://doi.org/10.1145/3589766

Case Study 1 - Energy-Efficient Approximate Edge Inference Systems

Approximate Computing

Approximate Systems

Current systems focus on how to use approximations to optimize the performance and energy consumption of *computing systems* that run DNN-based applications

- In most edge devices, the computation subsystem only contributes a modest amount to the total energy consumption of the system, which fundamentally limits the benefits of these approximation techniques

Approximate edge inference system (AxIS)

- Proposes a systematic methodology to perform joint approximations between different subsystems in a deep neural network(DNN)-based edge inference system
- Significant energy benefits compared to approximating individual subsystems in isolation

AxIS - Components

Hardware

- Intel Stratix IV GX-based Terasic TR4-230 FPGA development board
- The system was interfaced with a 5-MP CMOS image sensor as the sensor subsystem, 1-GB DRAM as the memory subsystem, and the ESP-WROOM-02 module as the communication subsystem. An Intel Nios II soft processor core was used as the compute subsystem

Classification

- Six large DNNs and four compact DNNs running image detection and classification applications
- Benchmarks
  - Classification: Four small DNNs - SqueezeNet1.1, MobileNetV2, MNASNet1.0, and Efficient-Net_Lite, and six large DNNs, namely AlexNet, VGG19_BN, DenseNet121, InceptionV3, ResNet101, and EfficientNet
  - Object detection: Faster_RCNN, Mask_RCNN, EfficientDet, and YOLOv5
  - Segmentation benchmark: Mask_RCNN

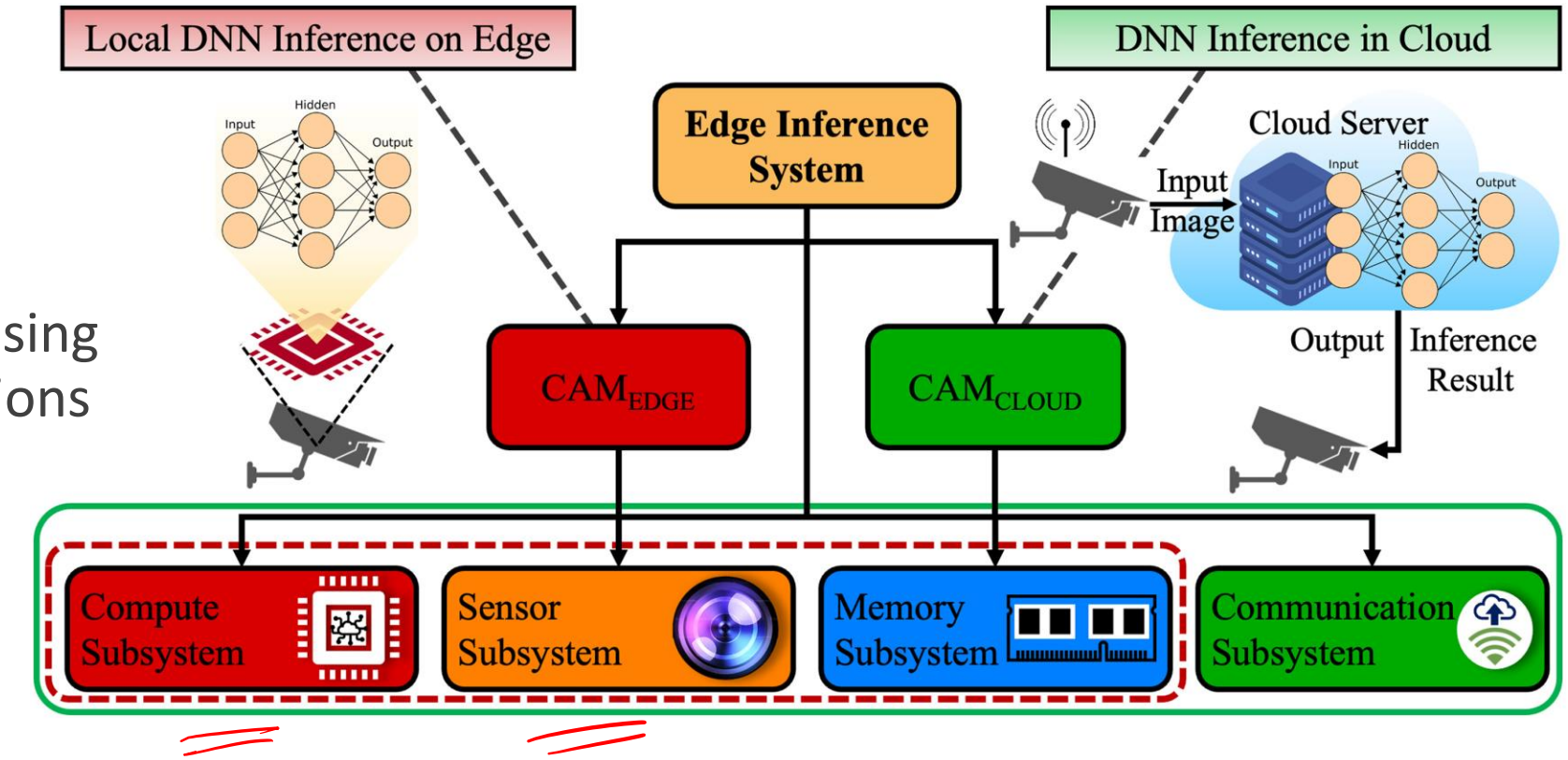
AxIS – Architectural Variants

Two variants: CamEdge and CamCloud

Subsystems: Sensor, Memory, Compute and Communication

AxIS performs DNN inference using three computer vision applications

- Image classification
- Object detection
- Image segmentation



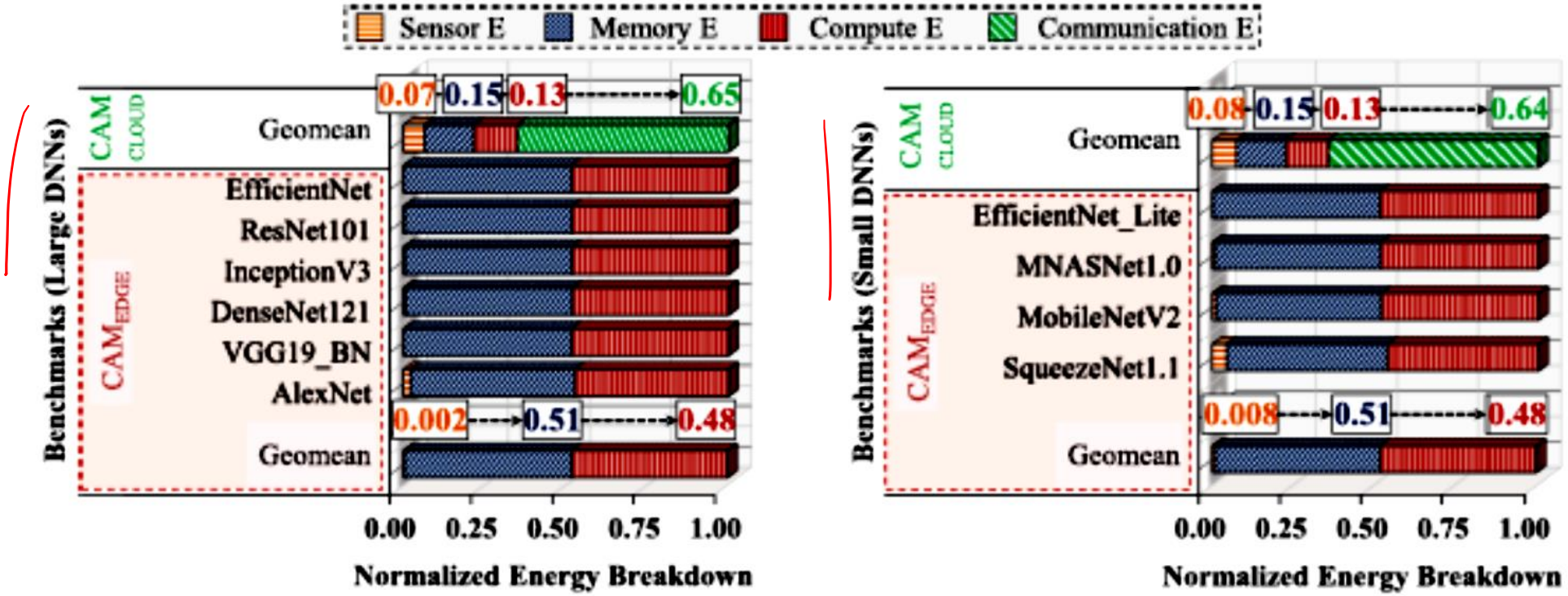
Discussion

What would be the difference between CamEdge and CamCloud wrt

- *Sensor energy*
  - *What factors does this depend on?*
- *Communication energy*
  - *How is this calculated?*
  - *What factors does this depend on?*
- *Compute energy*
  - *How is this calculated?*
  - *What are the factors that contribute to compute energy in CamEdge and CamCloud?*
- *Memory energy*
  - *What are the memory read/write operations that happen in the CamEdge architecture?*

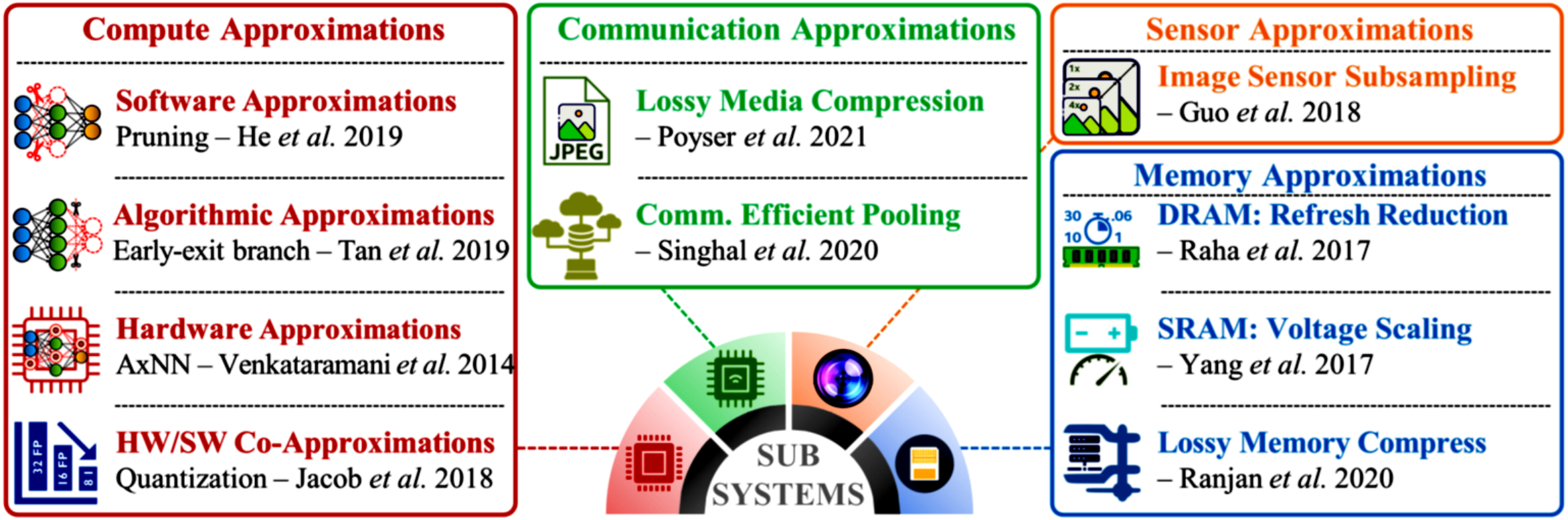
AxIS - Subsystem energy breakdown for image classification DNNs

Esys = Esens + Emem + Ecomp + Ecomm



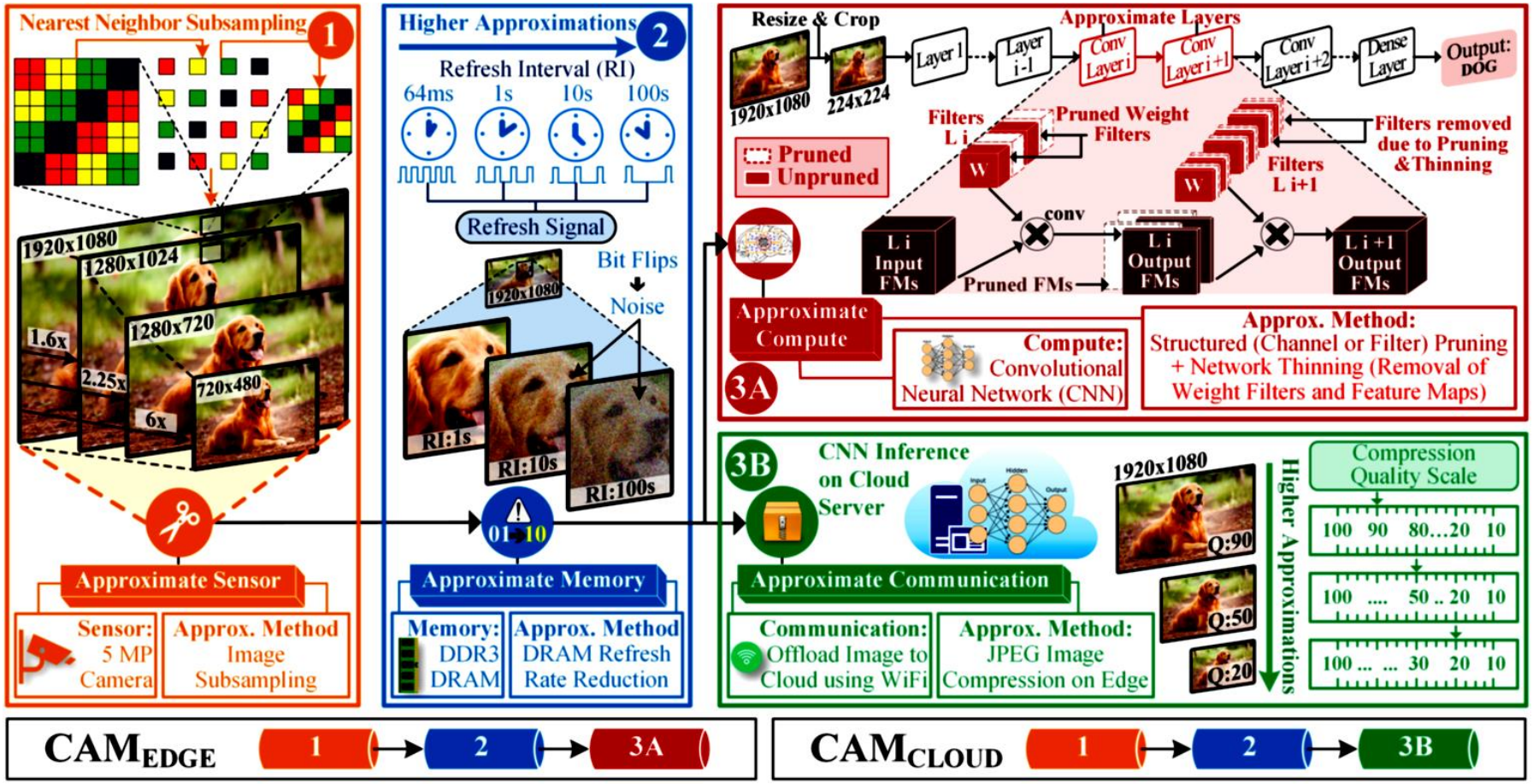
Subsystem energy breakdown for image classification DNNs

AxIS - Possible Approximations



Landscape of DNN-based approximations applied to different subsystems

AxIS - Subsystem-Level Approximations



Compute approximations in axis

Structured pruning and thinning to facilitate energy-efficient inference on resource-constrained edge devices

QUSP to approximate the DNN inference operation

- The system designer provides pretrained DNN model, a set of end user defined target application quality loss bounds (e.g., {0.5, 1.0, 2.5, . . .}), a DNN-specific application dataset, and a saliency metric for ranking structures (filters or channels) in a DNN
- QUSP generates a gradual/iterative quality-driven pruning schedule and subsequently runs the structured pruning and thinning algorithm for compression of the DNN model
- Given a set of quality metrics, this framework generates a family of compressed DNN models each meeting a distinct quality bound requirements
- Generates multiple models with an accuracy-FLOPs tradeoff that can be directly executed on COTS edge devices in an energy-efficient way, without the need for any hardware modifications

Further

How approximations in one subsystem impacts the other subsystems

Case Study 2 – Design of a Geriatric Healthcare System

Ref: Ramachandran, Anita. 2021. Design and Development of Intelligent Geriatric Healthcare System [Doctoral thesis]. https://shodhganga.inflibnet.ac.in/handle/10603/472823

Requirements

Monitor for falls in old people

- Parameters: IMU sensor data, heart rate

Wrist worn device

Long term data analysis

Deployment scenario: geriatric care home

Considerations

How often to collect data

Threshold based detection

- Issues – false positives

ML based detection

- How do we collect data?
- Where do we store collected data? On the device? On the cloud? Elsewhere?
- Where do we do inferencing? On the device? On the cloud? Elsewhere?
- Power considerations
  - Sensing power, Communication power, Processing power
- Latency requirements
- Performance metrics

Discussion - Possible Architectures

Collection

Preprocessing / feature engg

Xmit

Inferencing

Storage

Discussion - Qualitative Evaluation

	Device	Edge	Cloud	Memory	Sensing Power	Comm Power	Processing Power	Latency	Accuracy
1	Data collection No pre-processing Data transmission	x	Data storage Inferencing	L	C	H	L	H	H
2	Data collection Data pre-processing Data transmission	x	Data storage Inferencing	L+	C	H(H-)	L+	H-	H+
3	Data collection Data pre-processing Data transmission	Data aggregation Inferencing	Data storage	L+	C	M			
4	Data collection Data re-processing Data transmission Inferencing	Data aggregation	Data storage	H	C	L	H	L	L

Discussion - Qualitative Evaluation

Sl No	Device	Edge	Cloud	Memory	Sensing Power	Comm Power	Processing Power	Latency	Accuracy
1	Data collection No pre-processing Data transmission	x	Data storage Inferencing	L	=	H	L	H	H
2	Data collection Data pre-processing Data transmission	x	Data storage Inferencing	L	=	Lower than (1)	L	Maybe lower than (1) (Why?)	H (Why?)
3	Data collection Data pre-processing Data transmission	Data aggregation Inferencing	Data storage	L	=	Lower than (2)	L	Lower than (2)	H
4	Data collection Data re-processing Data transmission Inferencing (normal ML models)	Data aggregation	Data storage	H	=	L	H	L	H
5	Data collection Data re-processing Data transmission Inferencing w/model compression	Data aggregation	Data storage	Lower than (4)	=	L	H (but may lower than (4) (Why?)	L	Lower than (4)

Quantitative Evaluation - Considerations

What are our options for sensors?

What are our options for end device?

- Memory, support for low power modes of operation, peripherals compatible with sensors, support for network protocols, support for model compression techniques, if applicable

Profiling network performance

- How much data is generated per second per device?
  - (No: of parameters x no: of bits per parameters x no: of values sensed per second) + header size
- How many devices can be supported

Profiling energy consumption

Profiling memory usage

.. And thus derive the system level trade-offs

Other factors

Data security

Scale of deployment

More parameters

Backup

Approximations (1/4) – Compute

Compute approximations

- Algorithmic approximations
- Software-based approximation
- Hardware-based approximation
- Hardware-software co-approximation
- Several strategies for power optimization of the processor/CPU
  - Energy efficient CPU scheduling algorithms
  - Sleep modes of operations

Approximations (2/4) – Sensor

Sensor and Data Approximations

- Limited energy budget of IoT applications
- Approximations in the data acquisition path of such smart sensors
- Sensor subsampling, modulation of spatial/temporal resolution, quantization of sensor data
- Non-approximation based techniques
  - Compressive sensing, predictive coding, and DCT etc to optimize the ADCs, which are usually power and performance bottlenecks in high-resolution image sensors
- Require hardware modifications to COTS image sensors and to system applications

Approximations (3/4) – Memory

Memory and Storage Approximations

- Memory
  - Techniques can be classified in terms of the target memory type – SRAM (allocation of data bits), DRAM (DRAM refresh rates, write recovery time)
- Storage
  - Proposed in nonvolatile memory
  - Reduction of guard bandwidth, selective application of ECCs
- Also, non-approximation-based memory energy saving strategies
  - Require custom memory architectures and design that limits their widespread applicability for consumer devices

Approximations (4/4) – Communication

Communication approximations

- Lossy/lossless data compression
- Offloading of DL computations from edge devices to the cloud

innovate achieve lead

BITS Pilani

WILP

ML Systems Optimization

Shan Sundar Balasubramaniam

innovate achieve lead

BITS Pilani

WILP

ML Systems Optimization

• Shan Sundar Balasubramaniam

- Federated Learning

Federated Learning

- Scenario:
  - Data is available with multiple clients (devices, servers, or organizations)
  - A single (centralized) model is the requirement
  - Training can be done collaboratively
- i.e. learning task is to be solved by a federation of participating clients
  - and hence the term ***federated learning***

Federated Computing

- This notion is more general than ML:
  - Cryptographic approaches have existed for exchanging secret data and for centralized computing with encrypted data
  - Privacy-preserving data mining is about two decades old
  - Recently, federated approaches to social media networks are being adopted :
    - see Fediverse – community-owned, decentralized, privacy-centric social media networks
      - E.g. Mastodon (conforms to ActivityPub, a W3C standard)

Federated Averaging

- Federated Learning was first proposed by McMahan and Ramage
  - Learning from Gboards (Google Keyboard on Android)
  - Federated Averaging :
    - SGD run on local data on device
      - using a mini version of TensorFlow
    - Update communicated to the server
    - Server (samples,) averages and updates the central model

Federated Averaging - Algorithm

K clients η - learning rate E – number of local iterations Sample client size 0 <= C <= 1

```
// on each client k,   k = 1 to K;
clientUpdate(k,w) {
 batches = get_batches (dataset_k , batch_size)
 for i = 1 to E {
   for b in batches {
            grad = eval_gradient ( loss_function , b , w )
            w = w - learning_rate * grad
   }
 }
 return w
}
```

Contrast this with mini-batch SGD!

Federated Averaging - Algorithm

K clients η - learning rate E – number of local iterations Sample client size 0 <= C <= 1

```
// on each client k,   k = 1 to K;
clientUpdate(k,w) {
 batches = get_batches (dataset_k , batch_size)
 for i = 1 to E {
   for b in batches {
            grad = eval_gradient ( loss_function , b , w )
            w = w - learning_rate * grad
   }
 }
 return w
}
```

```
// n_j = |dataset_j|
//On the server:
Initialize w_0
for each round t = 1 to num_rounds {
   m = max (C*K, 1)
    S = select m clients at random
    for each client j in S in parallel {
        weights_t[j] = clientUpdate(j,w_{t-1})
    }
    tot_ex= Σ_{j∈S} n_j
    w_t = (1/tot_ex) * Σ_{j∈S}  (n_j * weights_t[j])
```

Contrast this with mini-batch SGD!

Characteristics of the setting

- Typical setting for Federated Learning:
  - Massively distributed
    - #clients > average training examples (per client)
  - Unbalanced
    - Varying amounts of local training data
  - Non-IID
    - Any particular user('s data) is NOT representative of the population distribution
  - Limited Communication
    - High Latency (including non-availability) and low bandwidth

Typical Training Process

- The typical process for Federated Training:
  1. Client Selection
     1. The Server samples a set of clients
        - e.g. mobile phones may ping the server only if they have a certain amount of power, idle, in a wi-fi-network, and so on.
  2. Broadcast
     1. The clients get the latest model weights and (possibly) a training program
        - Devices may not store all kinds of training programs or cannot decide
        - The program can be pre-compiled in some form (e.g. a TensorFlow graph)
  3. Client Computation
     - Each selected device computes an update to the model using the training program
  4. Aggregation
     - Server collects and aggregates updates
     - Synchronous aggregation – stragglers may be dropped
  5. Model Update

Federated Learning and Privacy

- One of the motivations for Federated Learning:
  - Compliance with data privacy requirements (and/or Governmental policies / regulations)
- Aggregation (see step-4 of the process in previous slide)
  - This step can be made secure (i.e. client updates are encrypted, anonymized etc.) to ensure privacy

Communication Limitations

- Client Selection can be tuned to include bandwidth (or latency) calculations to address
  - See Step-2 in the process (see slide 8)
- Client updates can be compressed
  - Lossy compression can be considered with appropriate trade-off in accuracy

Cross –device vs. Cross-silo Federated Learning

- In Cross-Silo Federated Learning,
  - data is held by multiple organizations (client) and training is locally done for each client
    - E.g. Medical, Financial, Data Centers in different regions
  - Typical number of clients:
    - 2 to 100
      - In contrast cross-device Federated Learning has to handle millions of devices (see next slide)

Scale of Cross-device Federated Learning

Parameter	Scale
Typical Population Size (# devices)	10^6 to 10^{10}
Typical Number of Devices Sampled in one Round	50 to 5000
Typical Number of Devices Participating in Training One Model	10^5 to 10^7
Typical Number of Rounds for Model Convergence	500 to 10000

Security Issues (beyond Data Privacy)

- Data Poisoning and/or Biasing the model are issues
- Solutions?
  - Authentication
  - Distributed cryptographic protocols for secure multiparty computation
    - Often requires special hardware
- Extreme case:
  - Byzantine-robust?

Systems for Federated Learning

- The primary requirement for a Federated Learning System is Scalability.
- Consider each phase/step of the process:
  1. Selection:
     1. Devices need to check-in with the server to be considered for selection
     2. Server may reject devices checking-in late.
  2. Broadcast:
     1. Server may need to read a **model-checkpoint** from stable storage for broadcast)
        1. Why?
  3. Devices may need Configuration before the Aggregation phase
     1. *Local Computation may simply be a lookup or selection from a store or sent by the server – this may require device initialization*
     2. *Secure Aggregation* or *Compressed communication* may need protocol initialization on the device.

Systems for Federated Learning

- Server has to be elastic:
  - It has to adapt to an increase/decrease in the number of devices participating/checking-in
  - Throughput, Response Time, and Utilization are metrics to be evaluated
    - Multi-threading
    - Multiple Servers
    - Edge Servers may act as aggregators for a group of devices
- Flow control
  - Server has to manage a large number of communications, check-in or update.

Systems for Federated Learning

- Multi-tenancy
  - in the form of multiple trainings on different population subsets
- may be a requirement
- Introducing intermediate servers (like edge servers)
  - will enable the master server to be elastic even in the presence of multi-tenancy.
- In the presence of a large number of devices and multi-tenancy,
  - coordinating and scheduling becomes the critical function of the master server:
    - For instance:
      - synchronizing between rounds, saving checkpoints etc.

Failures and Failure Handling

- With a large population,
  - significant percentage of devices may fail.
- Failure handling protocols have to be in place:
  - The simplest option is for the sever to ignore - in the short term –
    - and bring them in off-line.
- Failed node vs. malicious node?
  - Sybil attacks?
  - Byzantine failures?
  - Large scale failures?