Center of Excellence: Supercomputing for AI & Big-Data

Introduction to High Performance Computing and its impact on life

by: Tassadaq Hussain

Director Centre for AI and BigData

Professor Department of Electrical Engineering

Namal University Mianwali

Collaborations:

Barcelona Supercomputing Center, Spain

European Network on High Performance and Embedded Architecture and Compilation

Pakistan Supercomputing Center

Introduction

Education:

PhD. Barcelona-Tech
Microsoft Research, Infineon Technologies France, Microsoft Research Cambridge, IBM

Suspenseful record of academic management as Professor and Dean

Enhanced Education Quality by Inculcating Outcome Based Education by Applied and Sustainable Projects

Experience:

20+ year's versatile experience in the area of Computer Architecture, AI, Software Architecture, Big-Data Architecture
Served National and International Academia, Industry and Government

- **Barcelona Science Park Spain**
- **Cambridge Science Park UK**
- **Technopolis Of Sofia-Antipolis, France**

Innovation, Research and Commercialization

Innovation and Research

- **110+ Million Pkr National and Int'l Funding.**
  - Supercomputing and Artificial Intelligence
  - Smart Electric Motor Controllers
  - Biomedical Applications
- **100+ Publications**
- **10 Patents**
- **10 MVPs**
- **5 Int'l Collaborations**

Development & Commercialization

60+ Million of Industrial Investments.
Developed Digital Systems for Industry.
Transform Idea into product.
Innovation and Commercialization for Sustainable economic and industrial development.

Capacity Building:

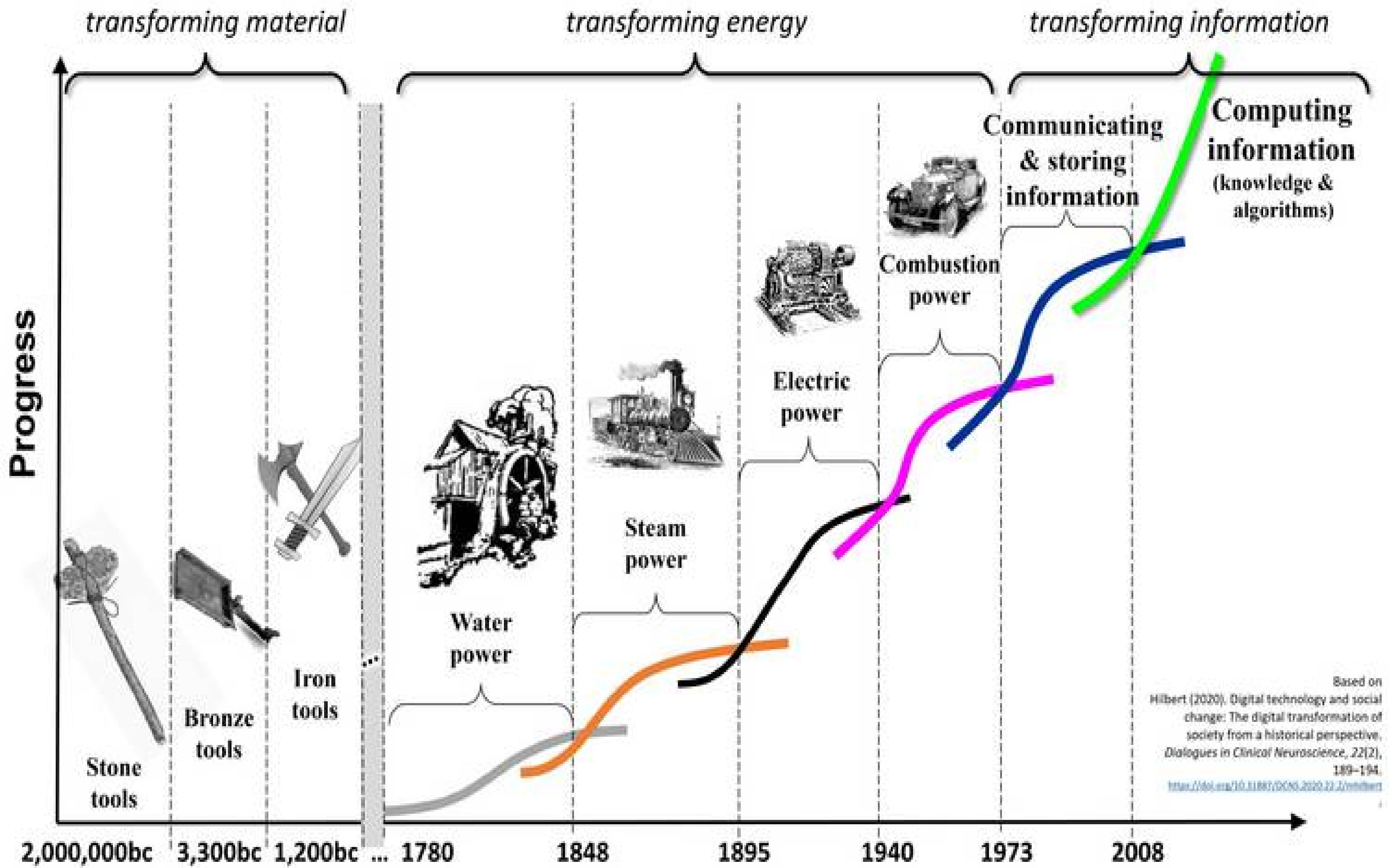
Conducted more than **50 national and international workshops** and **training on Commercializable research, Writing successful grant proposal, and research and innovation.**

Provides Consultancy and Support for Entrepreneurship, Start-ups, Business Innovation and Technology transfer.

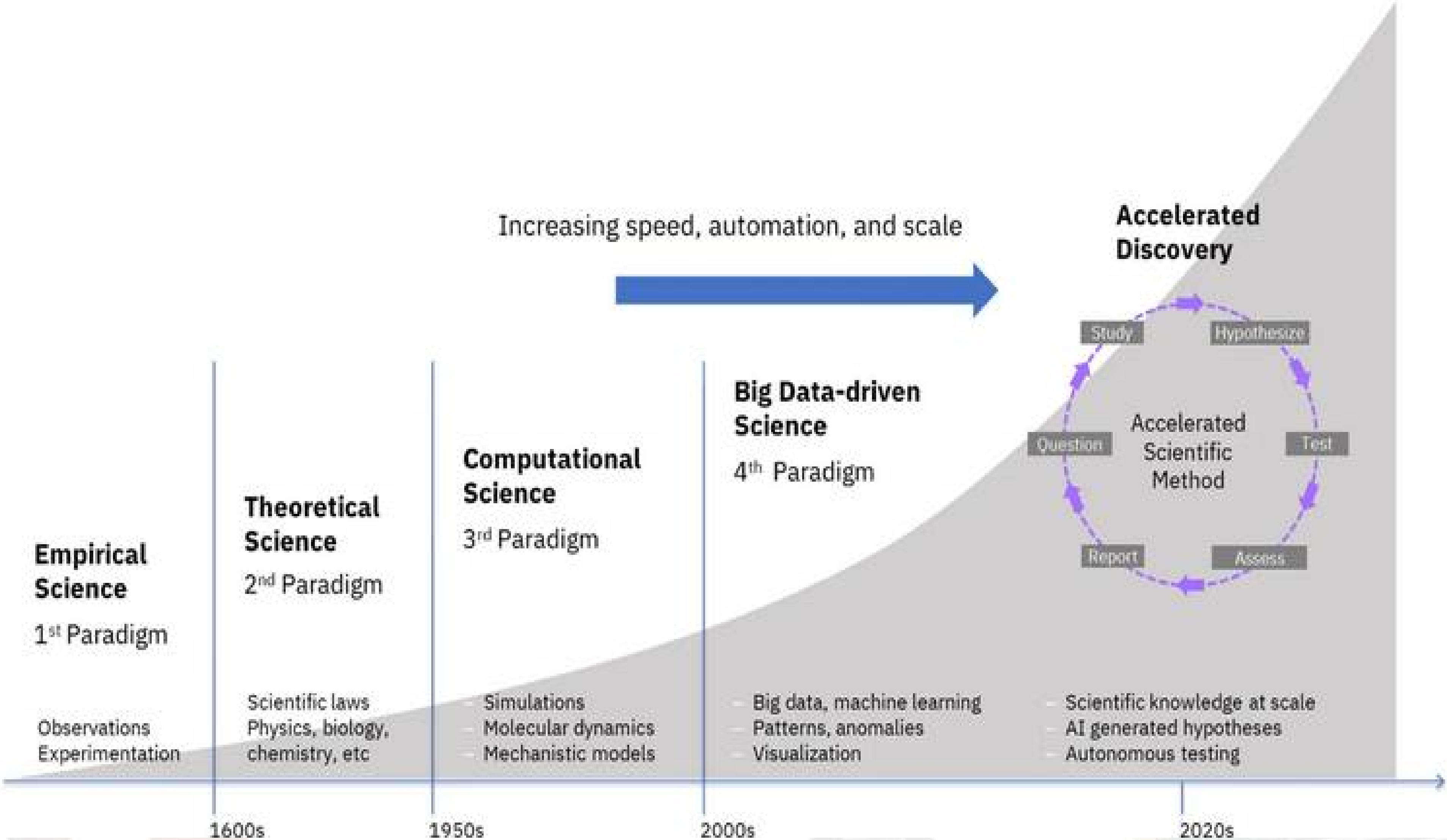
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
UVHC
Université de Valenciennes et du Hainaut - Cambrésis
New Zealand College of Chiropractic
graduating hands, hearts & minds
Pakistan Supercomputing
EntrepreneurialPark
Foot Analytic
ComputingPark

- **Mankind’s Progress & Revolutions**
- The Age of Big Data and AI
- The Role of High-Performance Computing (HPC)
- Namal Knowledge City & Supercomputing Facility

Human Progress: From the Stone Age to the Data Age

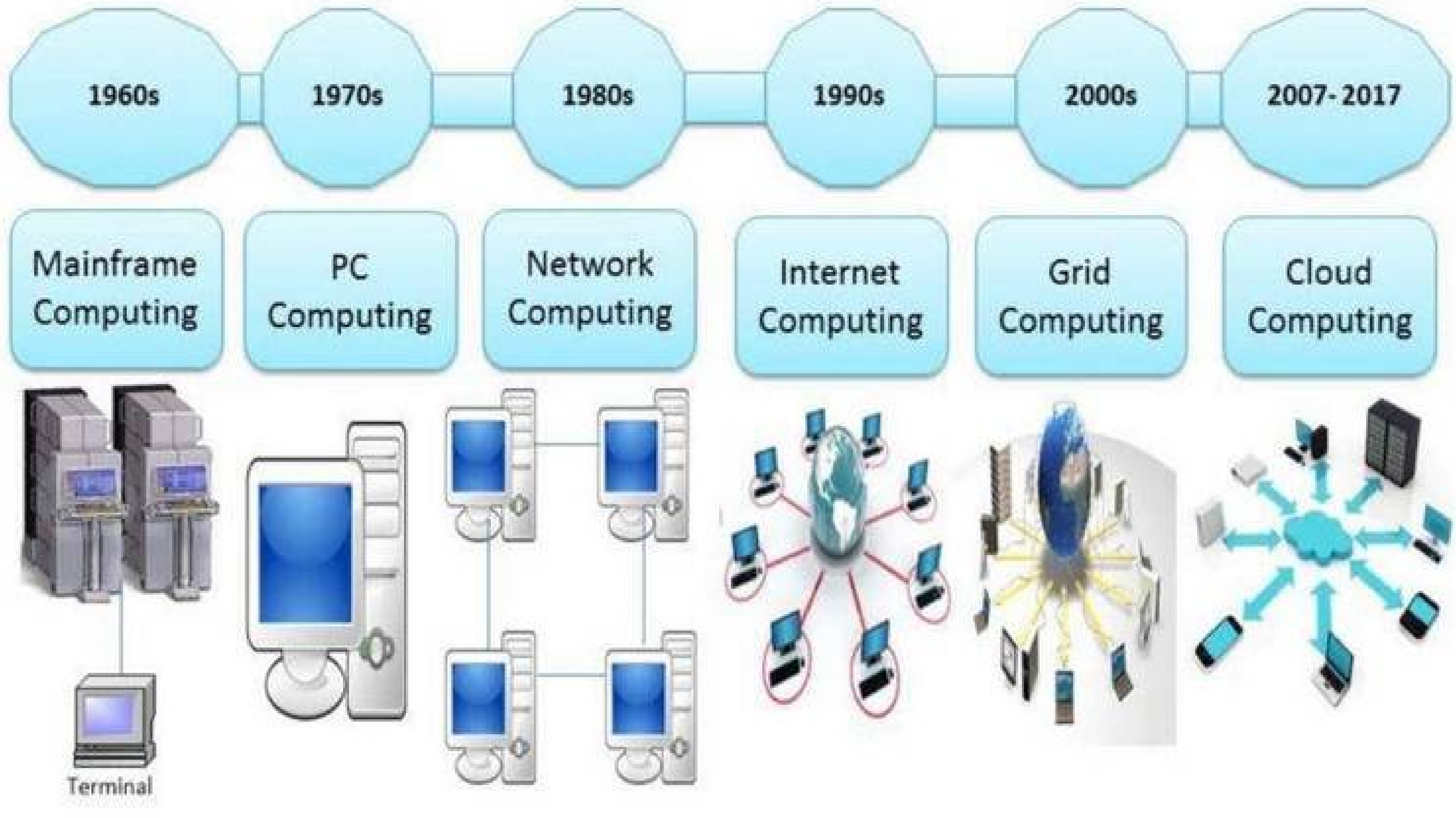


From Age of Empirical Science to Computational-Science

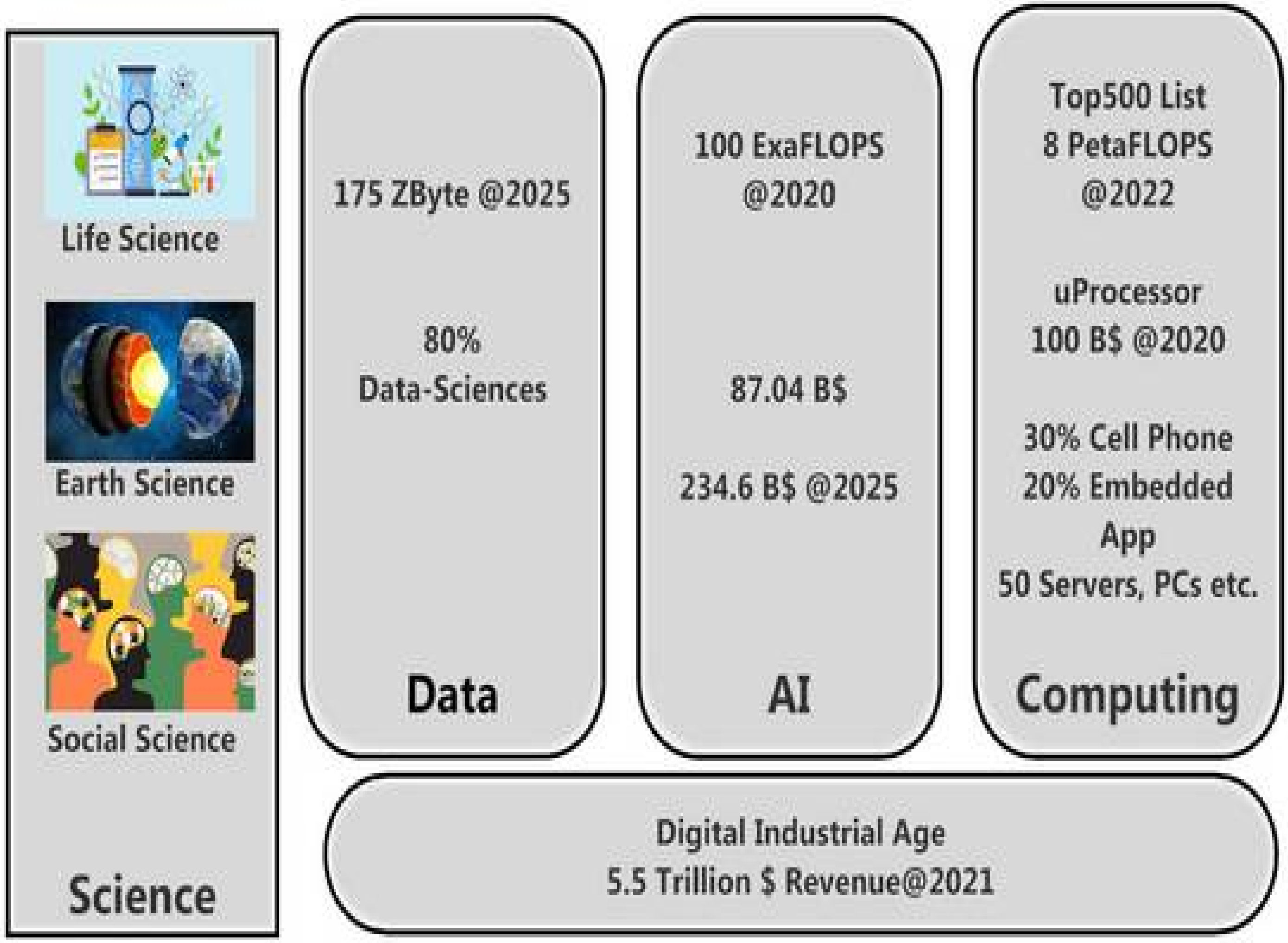


https://www.nature.com/articles/s41524-022-00765-z/figures/1

From Mainframe to Cloud Computing

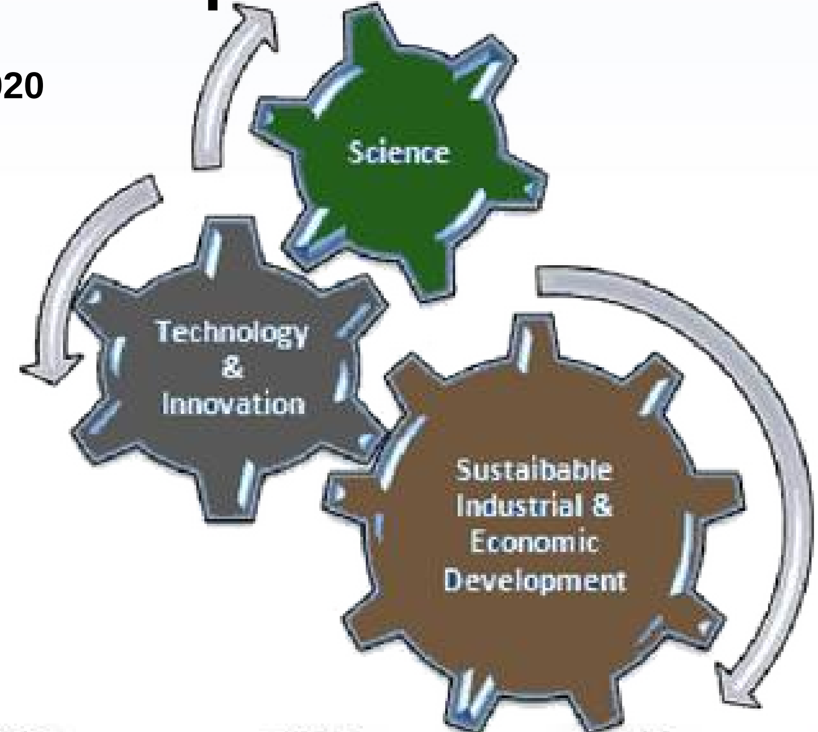


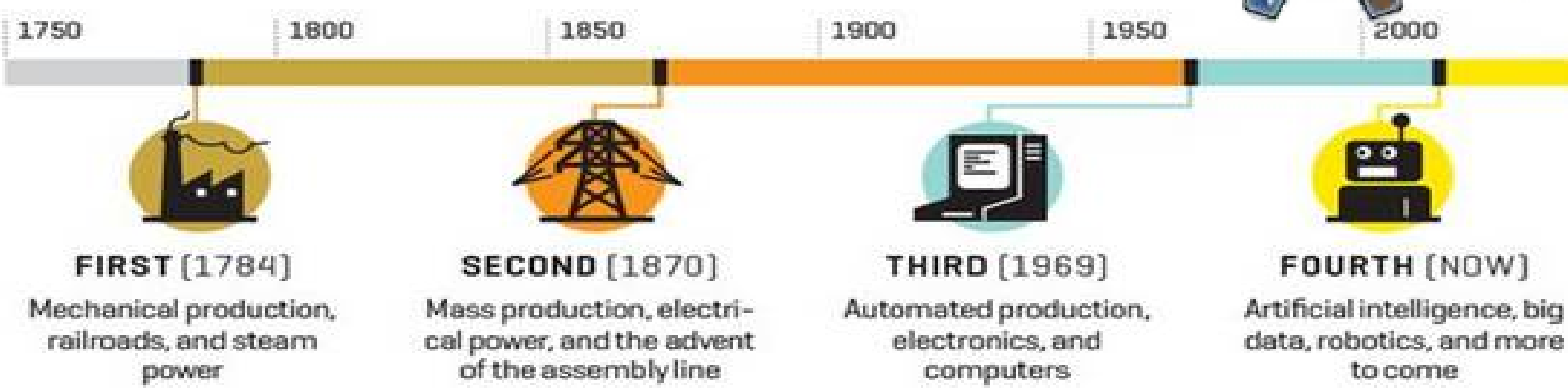
Ecosystem of Modern Industry



Industrial Revolutions and Sustainable Developments

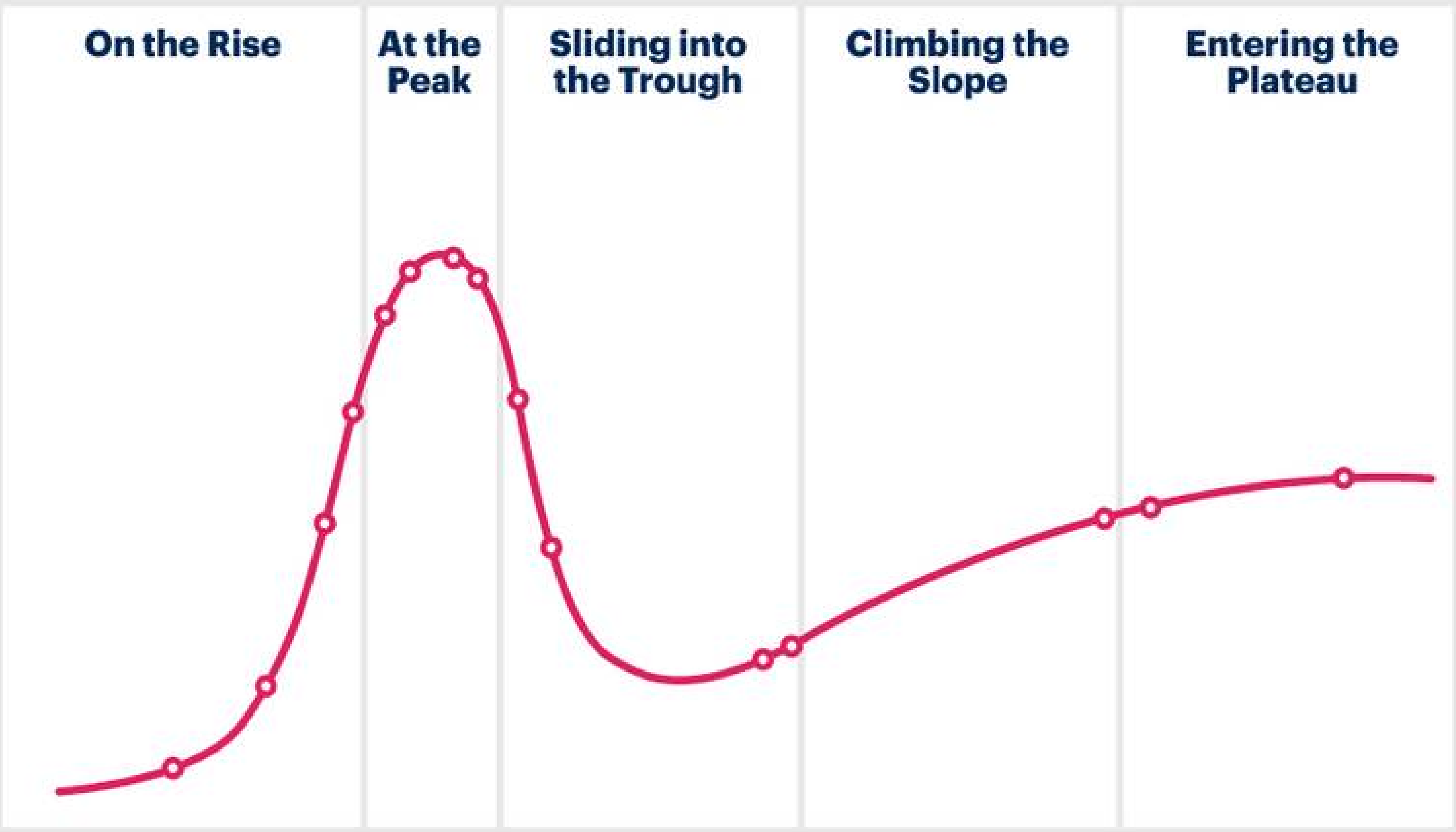
Top 10 Biggest Companies By Market Size From 2010 - 2020



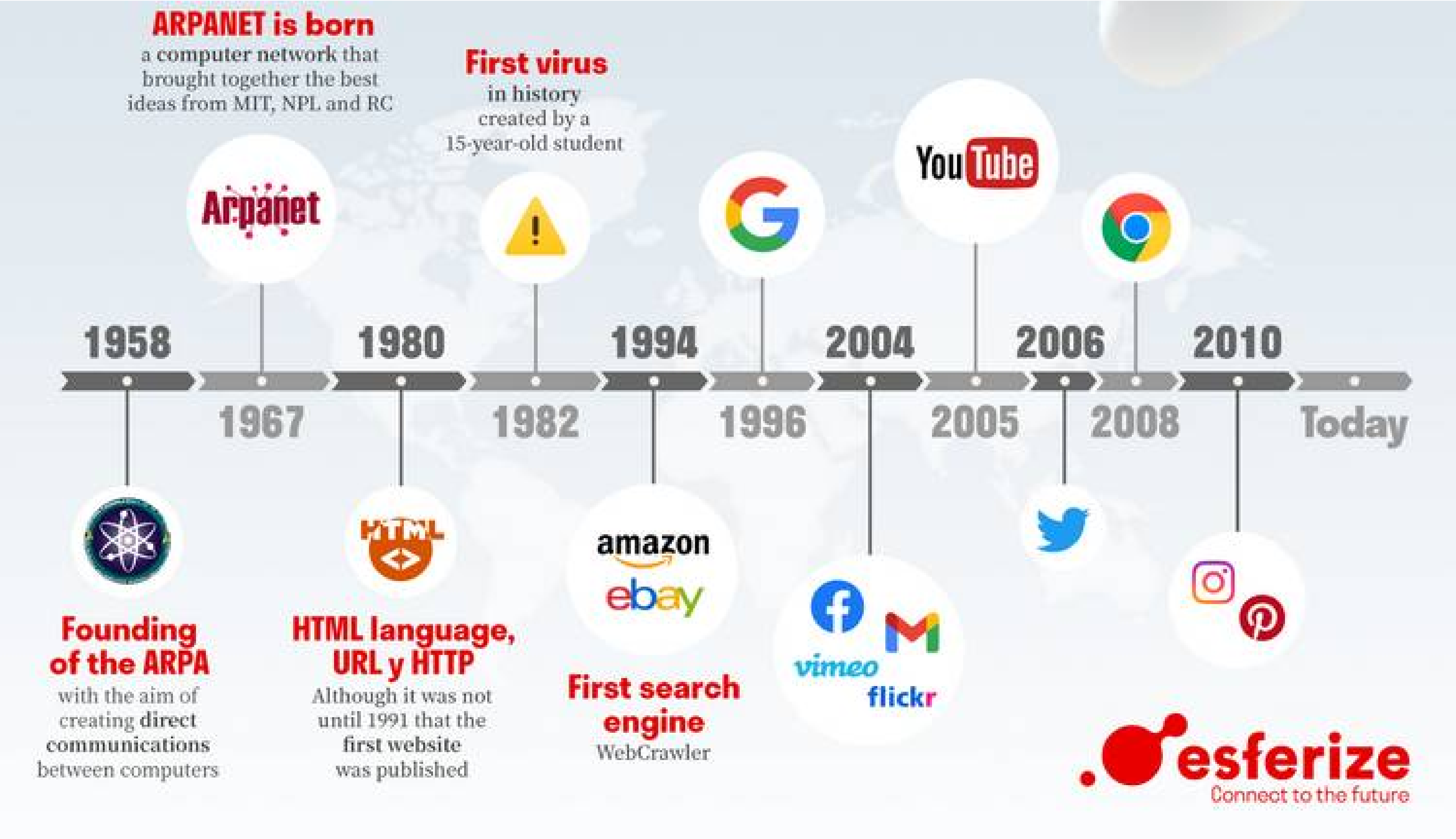


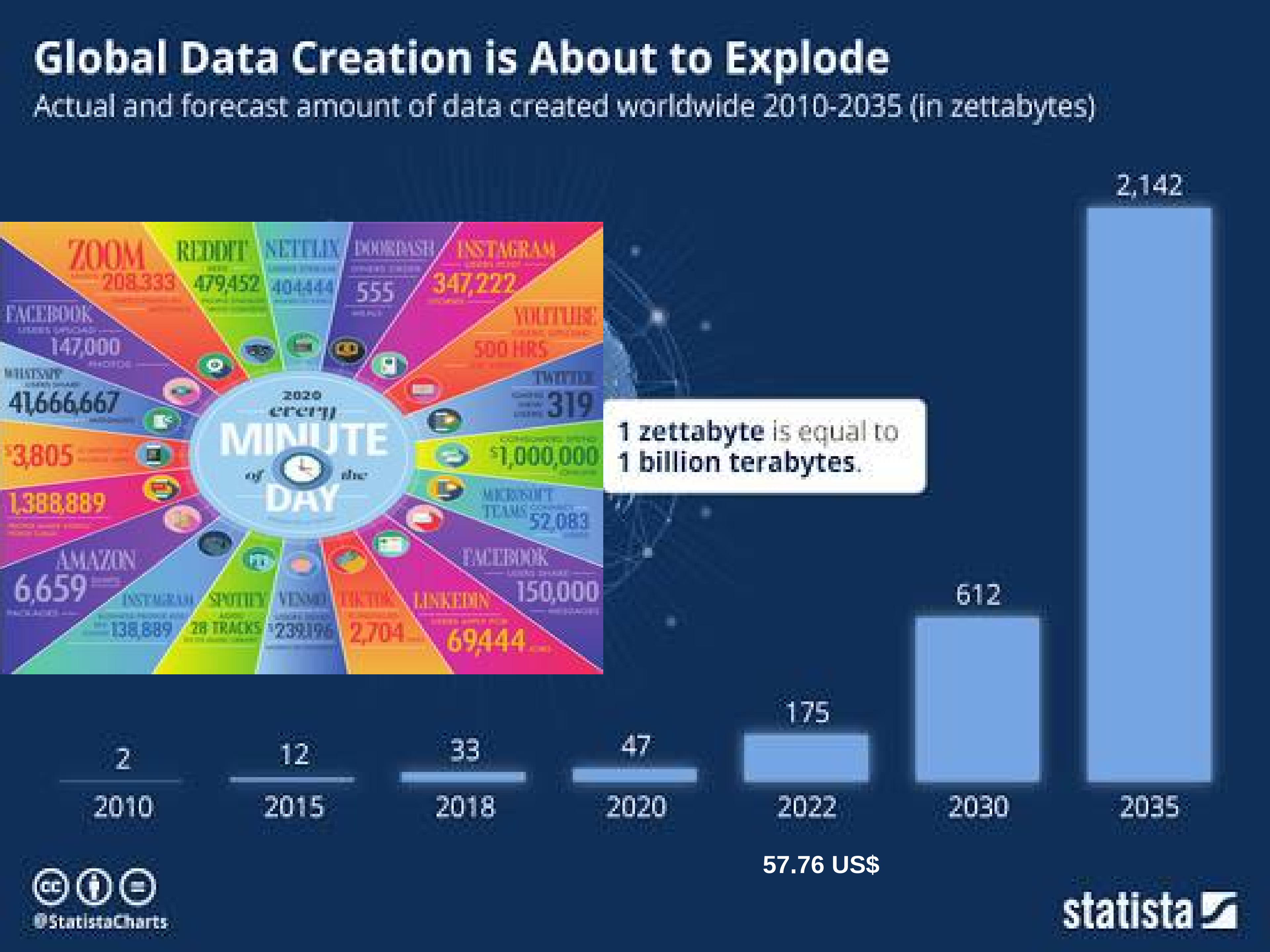
- Mankind’s Progress & Revolutions
- **The Age of Big Data and AI**
- The Role of High-Performance Computing (HPC)
- Namal Knowledge City & Supercomputing Facility

Understand Hype Cycle

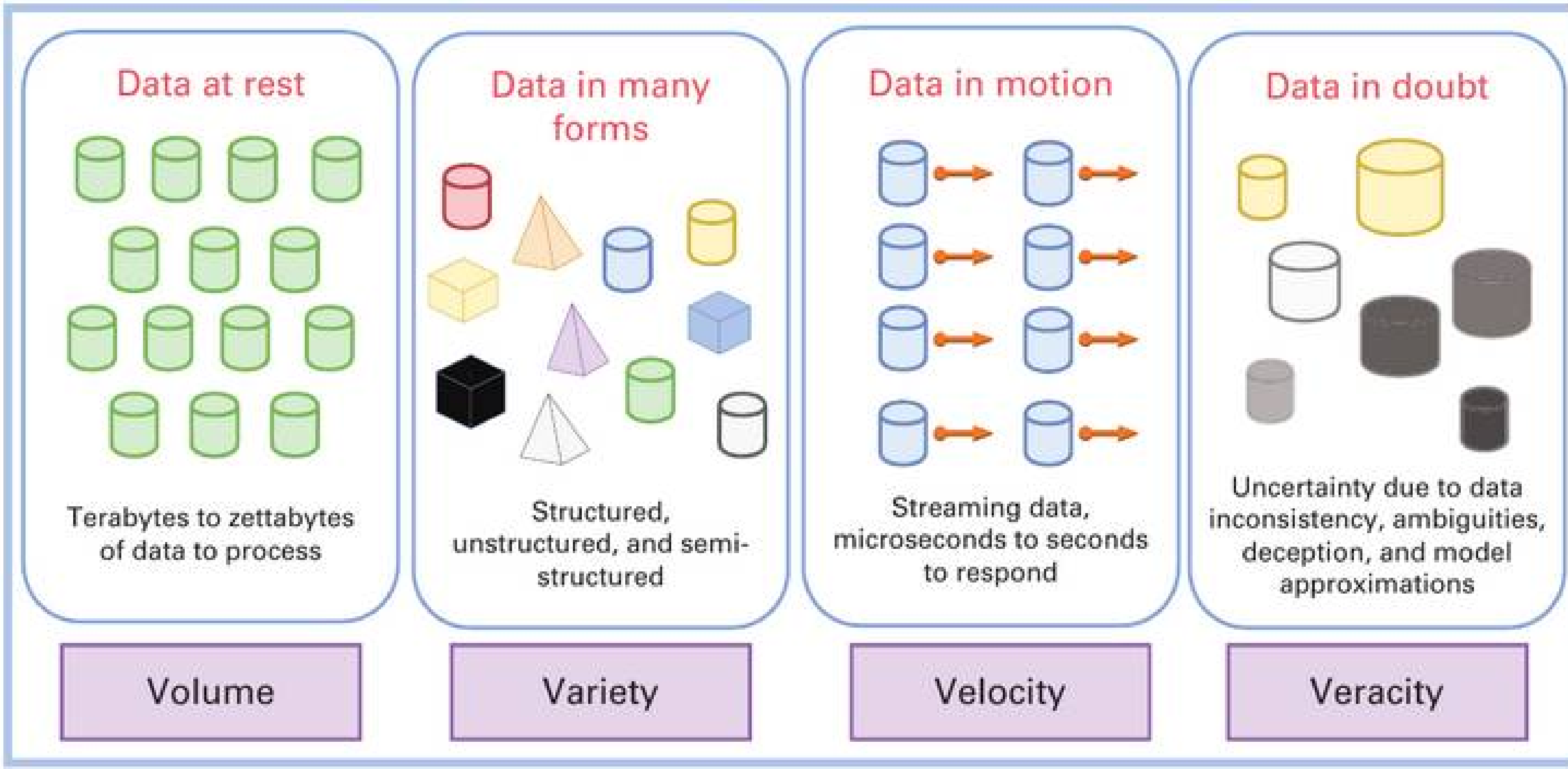


Evolution of Connectivity



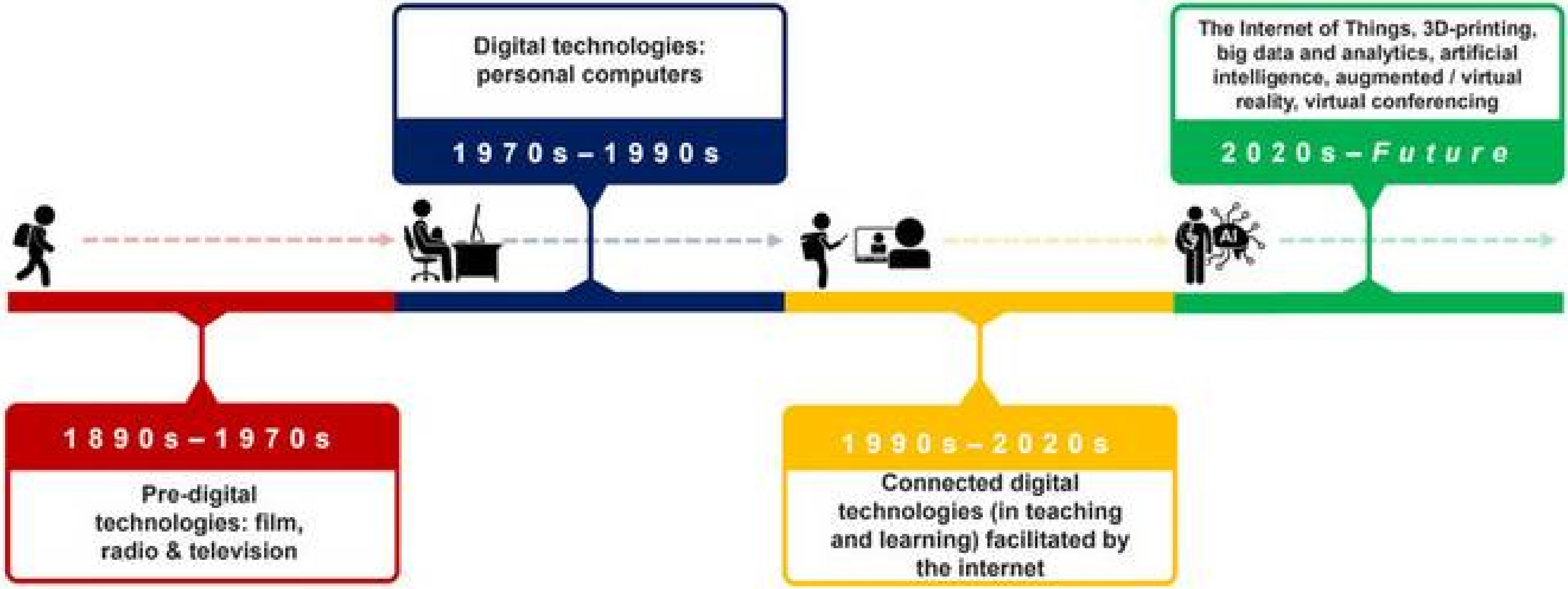

Global Data Creation is About to Explode
Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)
2020
every
MINUTE
of the
DAY
ZOOM
208,333
REDDIT
479,452
NETFLIX
404444
DOORDASH
555
INSTAGRAM
347,222
FACEBOOK
147,000
YOUTUBE
500 HRS
WHATSAPP
41,666,667
TWITTER
319
$3,805
$1,000,000
1,388,889
MICROSOFT TEAMS
52,083
AMAZON
6,659
FACEBOOK
150,000
INSTAGRAM
138,889
SPOTIFY
28 TRACKS
VENMO
$239,196
TIKTOK
2,704
LINKEDIN
69,444
1 zettabyte is equal to
1 billion terabytes.
2
12
33
47
175
612
2,142
2010
2015
2018
2020
2022
2030
2035
57.76 US$
@StatistaCharts
statista

Types of Data and its Challenges

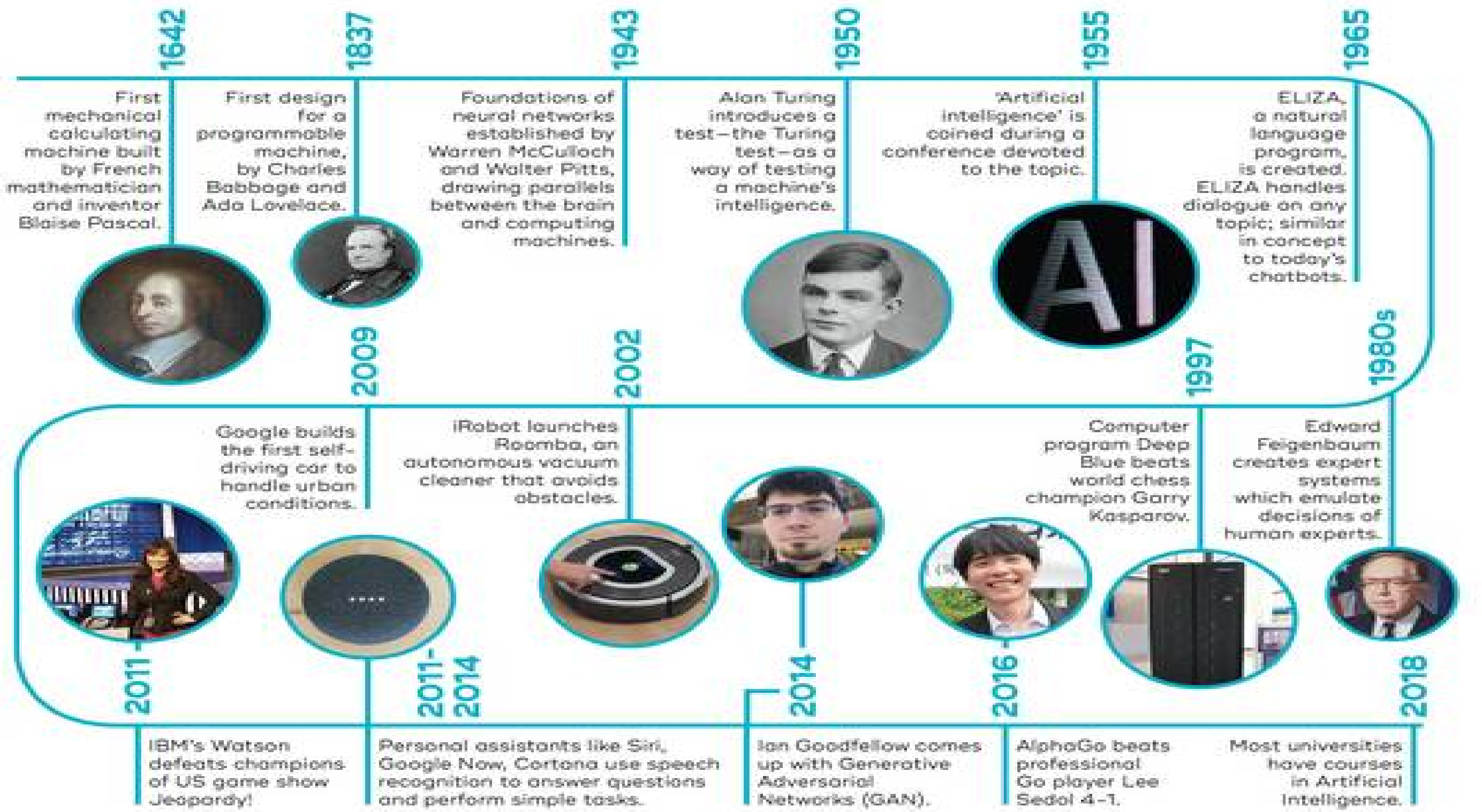


From Computing to Intelligence: The Evolution of the Digital Era

Era of Computing → Era of Data → Era of Information → Era of Intelligence



AI: The Only Solution for BigData



Past Present and Future

AGI

Artificial General Intelligence

ANI

Artificial Narrow Intelligence

Information/Big Data

Complex Adaptive Algorithms

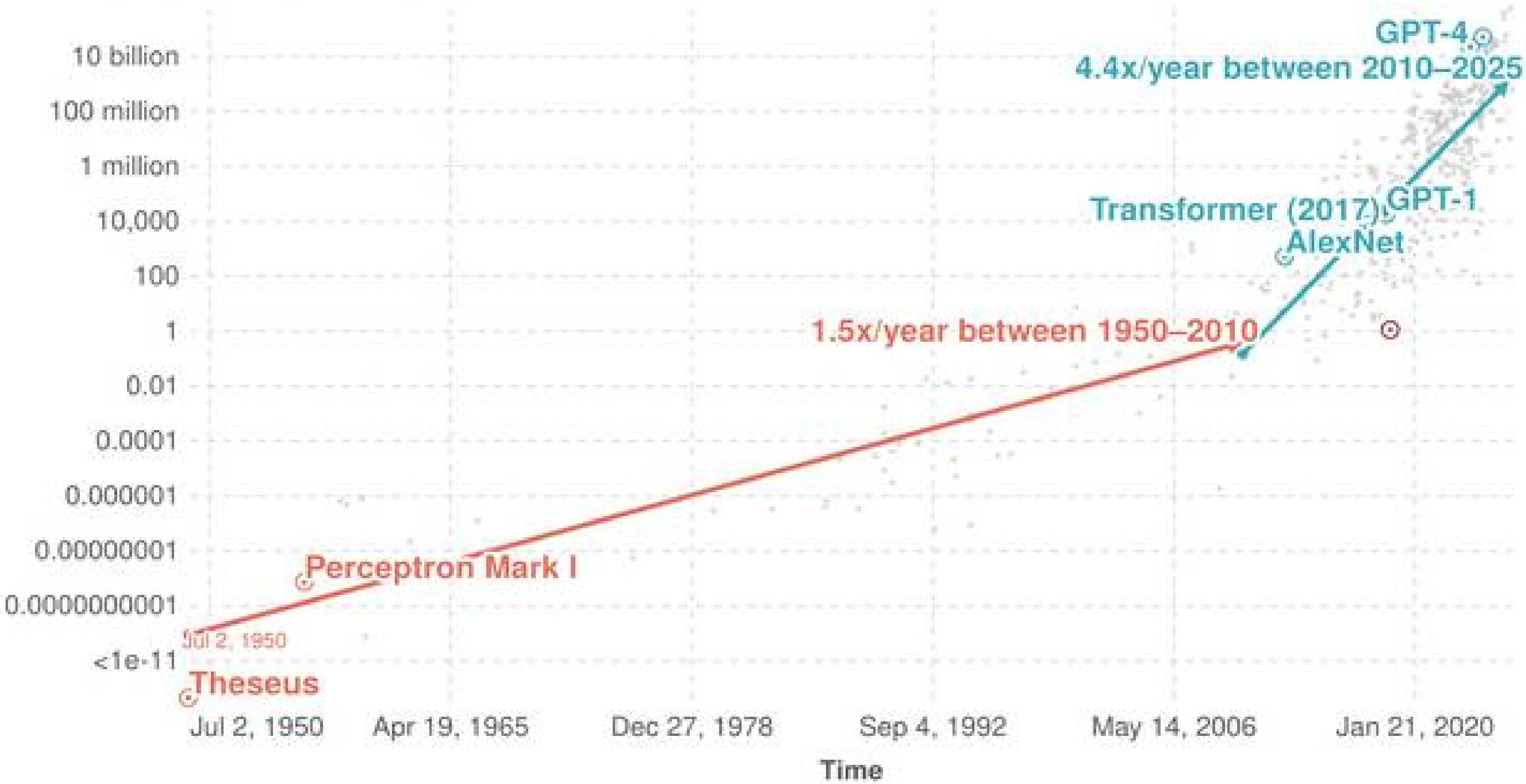
Computing Resources

BigData and AI Algorithms

- **Performance**
  - **Execution Time**
  - **Accuracy** “The accuracy of the model is inherently tied to the quality, diversity, and representativeness of the data used for training and evaluation."
  - **Scalability** “Methods that scale with computation are the future of Artificial Intelligence” — Rich Sutton,

Exponential growth of computation in the training of notable AI systems

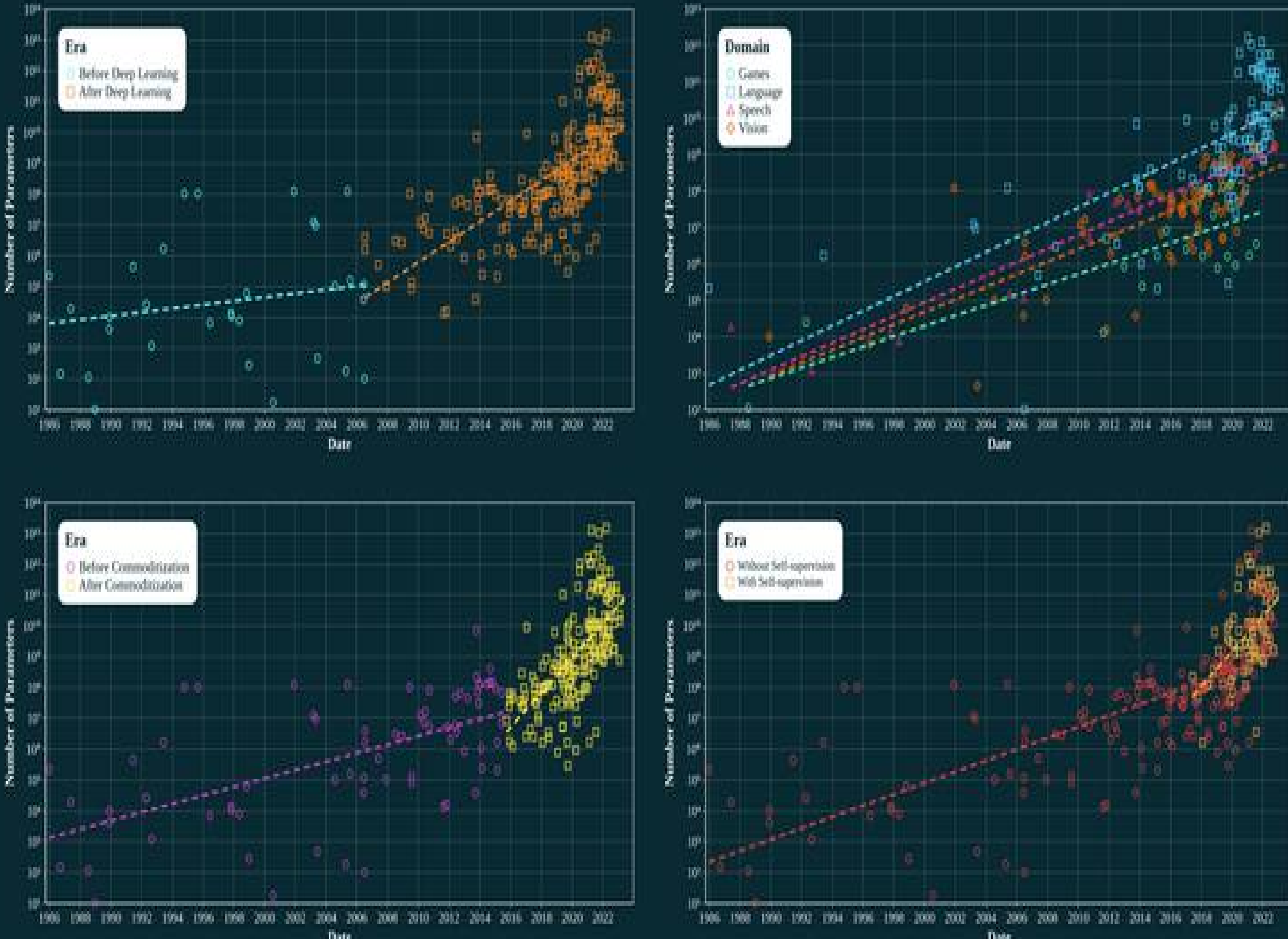
Computation is measured in total petaFLOP, which is 10^{15} floating-point operations[1].



Data source: Epoch (2025)

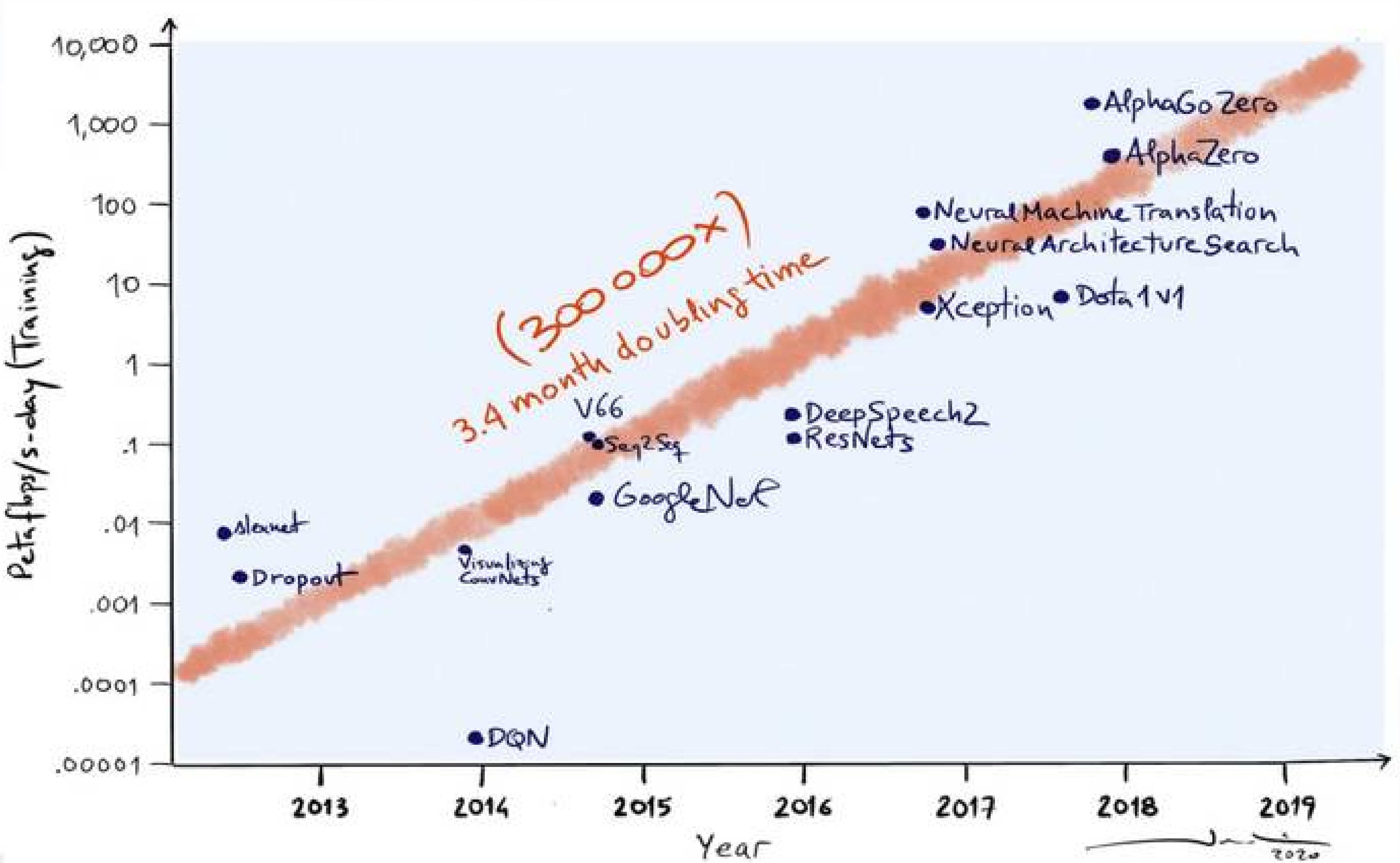
Note: Estimated from AI literature, accurate within a factor of 2, or 5 for recent models like GPT-4. The regression lines show a sharp rise in computation since 2010, driven by the success of deep learning methods that leverage neural networks and massive datasets.

1. **Floating-point operation** A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

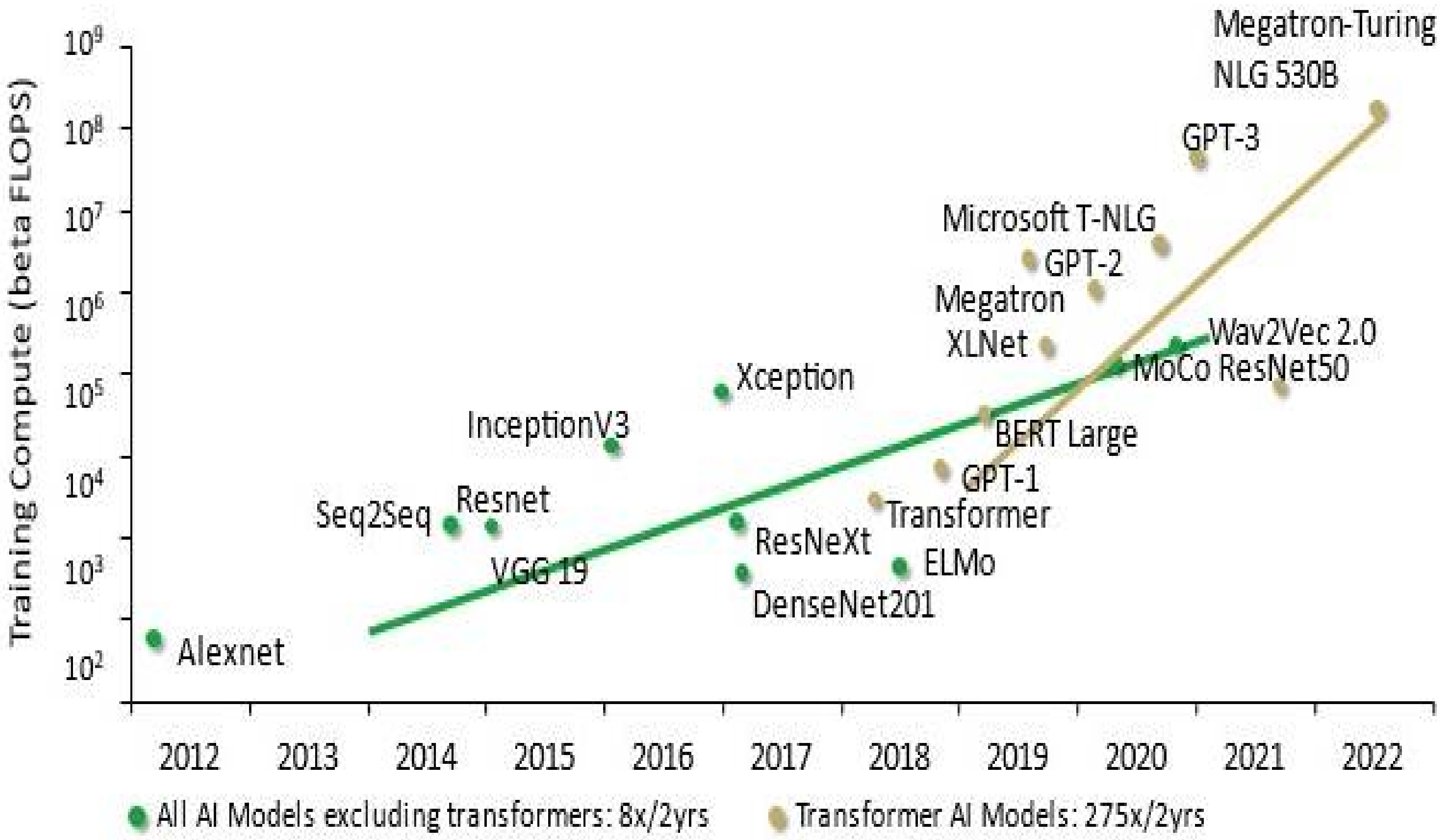


Era
Before Deep Learning
After Deep Learning
Domain
Games
Language
Speech
Vision
Era
Before Commoditization
After Commoditization
Era
Without Self-supervision
With Self-supervision
Number of Parameters
Date
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022

AI Computational Requirements



https://towardsdatascience.com/artificial-intelligence-is-a-supercomputing-problem-4b0edbc2888d


Training Compute (beta FLOPS)
10^2
10^3
10^4
10^5
10^6
10^7
10^8
10^9
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Alexnet
VGG 19
Seq2Seq
Resnet
InceptionV3
Xception
ResNeXt
DenseNet201
ELMo
Transformer
GPT-1
BERT Large
XLNet
Megatron
GPT-2
Microsoft T-NLG
GPT-3
Wav2Vec 2.0
MoCo ResNet50
Megatron-Turing NLG 530B
All AI Models excluding transformers: 8x/2yrs
Transformer AI Models: 275x/2yrs

- Mankind’s Progress & Revolutions
- The Age of Big Data and AI
- **The Role of High-Performance Computing (HPC)**
- Namal Knowledge City & Supercomputing Facility

High Performance Computing

- **Basics**
- Hardware Stack
- Software Stack
- Programming Model

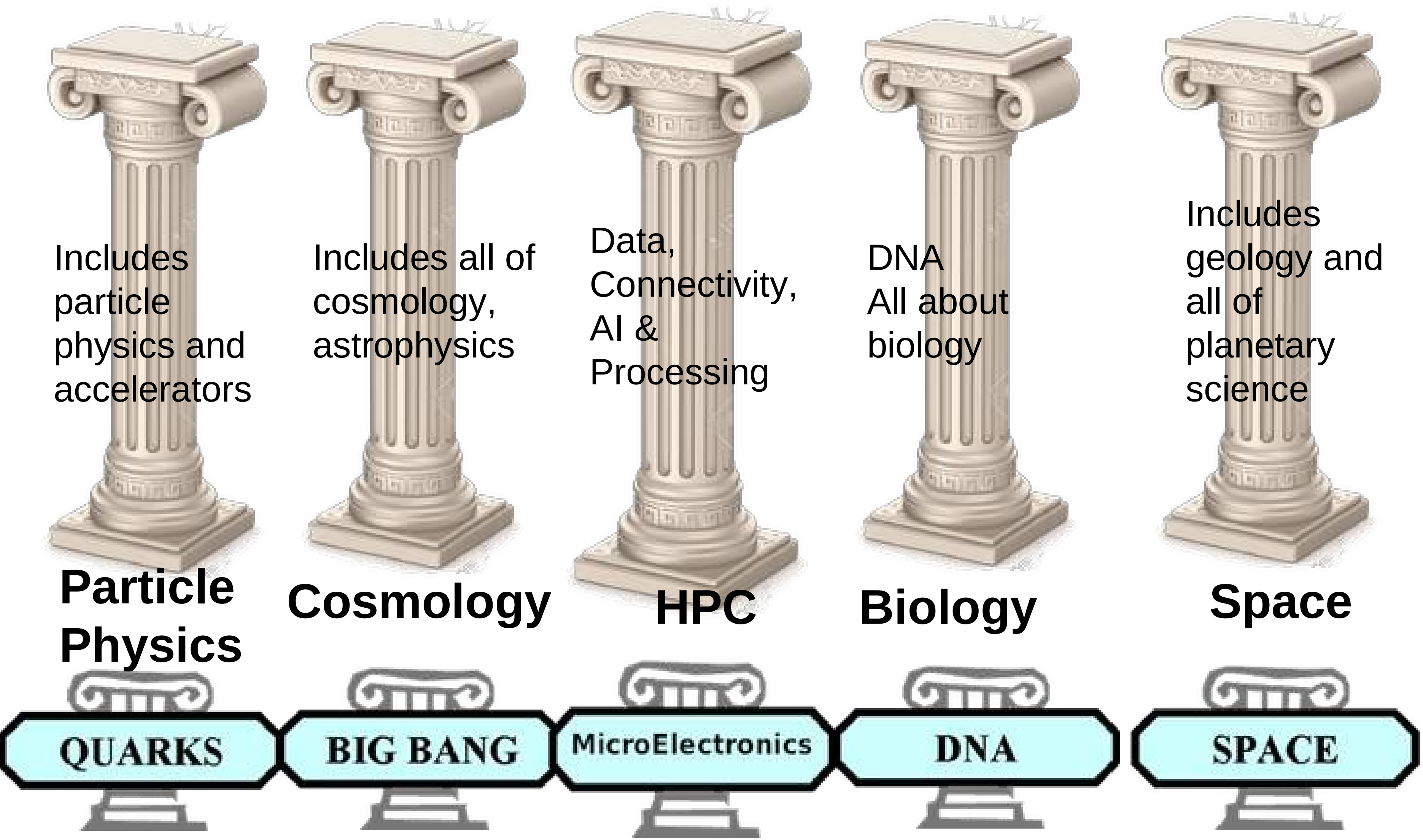
High Performance Computing?

HPC perform Scientific discovery, engineering simulations, AI/ML, and research that require massive numerical computation.

- Parallel processing across thousands of CPUs/GPUs.
- Optimized for floating-point operations per second (FLOPS), not just transactions.
- Used for simulations (climate, nuclear physics, genomics, aerospace), AI model training, and big data analytics.

Pillars of Science

Fermi National
Accelerator Laboratory



Criteria of HPC (in FLOPS)

Traditional Definition (1990s–2000s):

- HPC was defined as performance beyond what a typical desktop/workstation could achieve.
- Roughly > 1 GFLOP (10^9 FLOPS) was considered HPC in the 1990s.

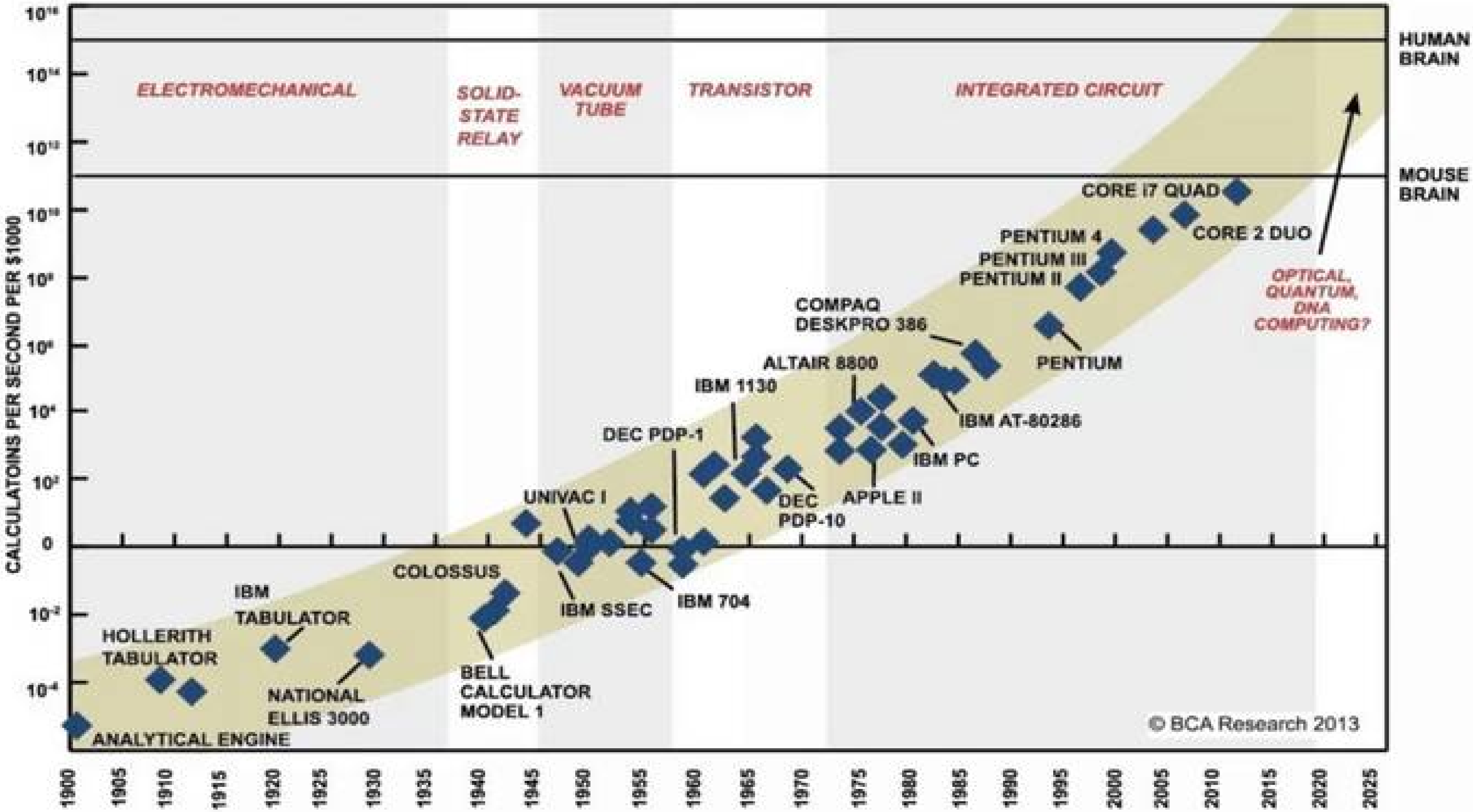
Cluster Era (~2000s):

- Systems with sustained > 1 TFLOP (10^{12} FLOPS) entered the HPC category.
- Example: In 2000, IBM’s ASCI White supercomputer reached 12.3 TFLOPS.

Modern Minimum (2020s):

- Today, a single GPU (e.g., NVIDIA A100) can exceed 10–20 TFLOPS, so the bar is much higher.
- For a cluster/system to be considered HPC, it generally should sustain at least 1 PFLOP (10^{15} FLOPS).
- Example: The TOP500 list (global supercomputer ranking) includes systems above ~1.2 PFLOPS (as of 2025).

Cost Vs Performance: Electromechanical to ICs

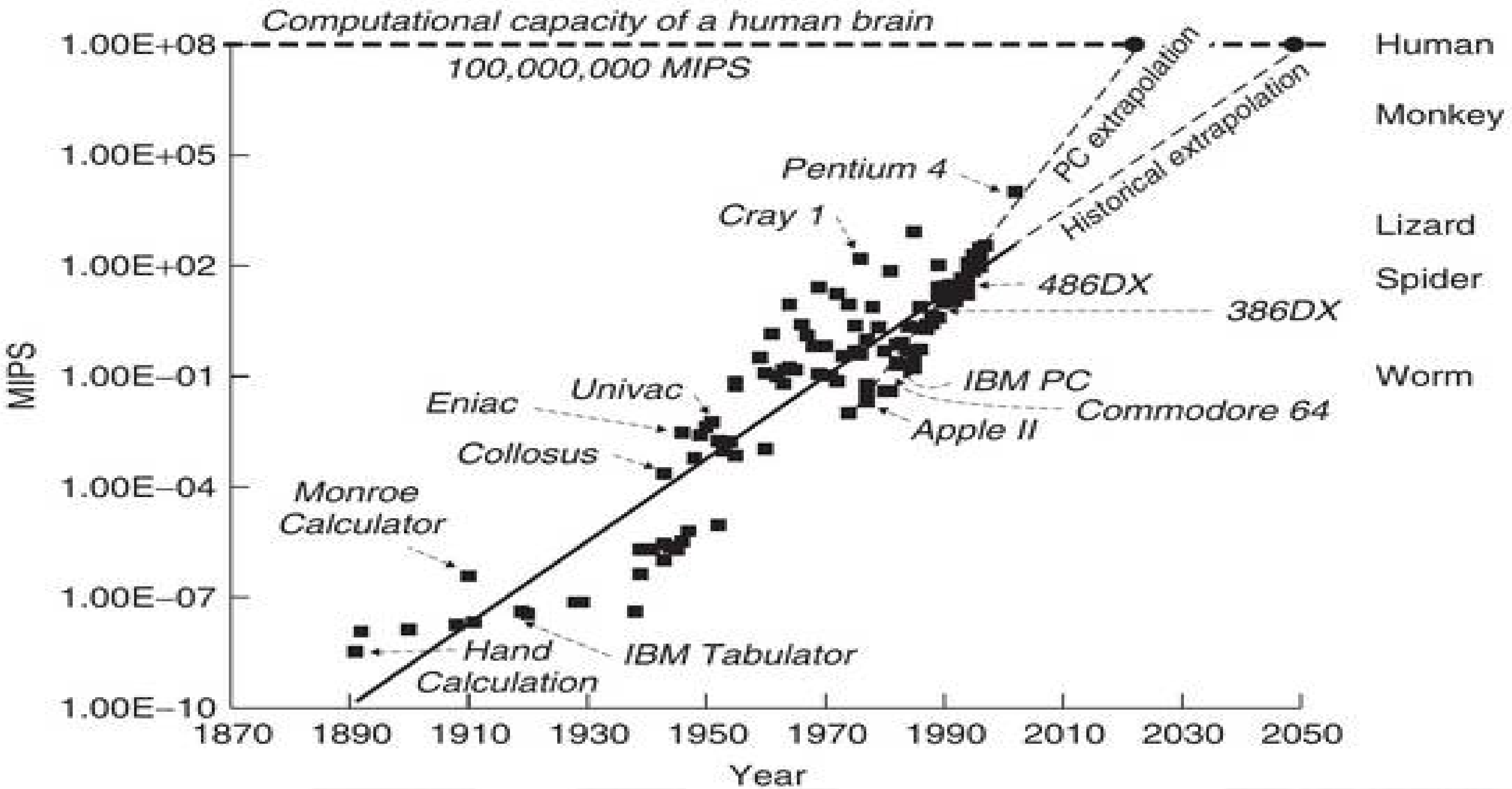


SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

1 Operation / Second = 1 B$

1B Operation / Second < 1$

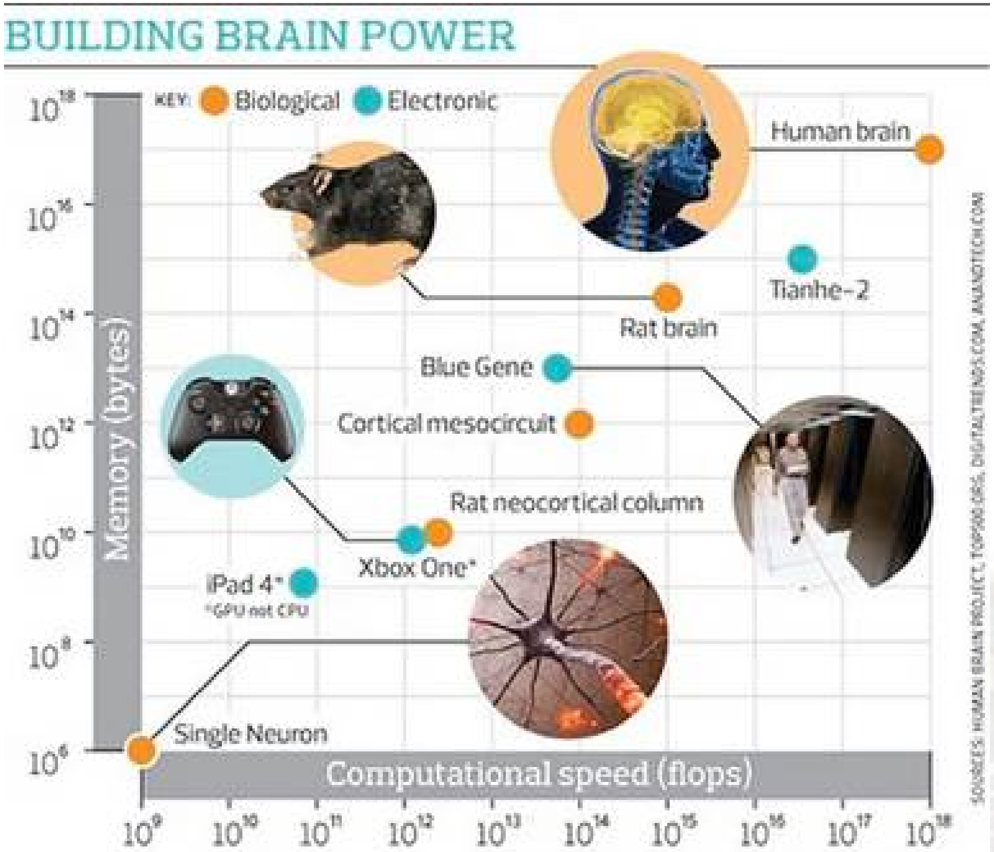
Human Computational Capabalities and HPC



It is estimated that sometime between the years **2025 and 2050**, a **personal computers** will exceed the calculation power of a human brain.

Compute Vs Intellectual Capability

- Perform around 1 exaFLOP (10^{18} FLOP/s) with just ~20 watts of power.
- The adult human brain's memory capacity is often estimated at approximately ~2.5 petabytes.
- Aurora Supercoputer gives 1.012 exaFLOP (Rmax) @ ~38.7 MW



https://www.theguardian.com

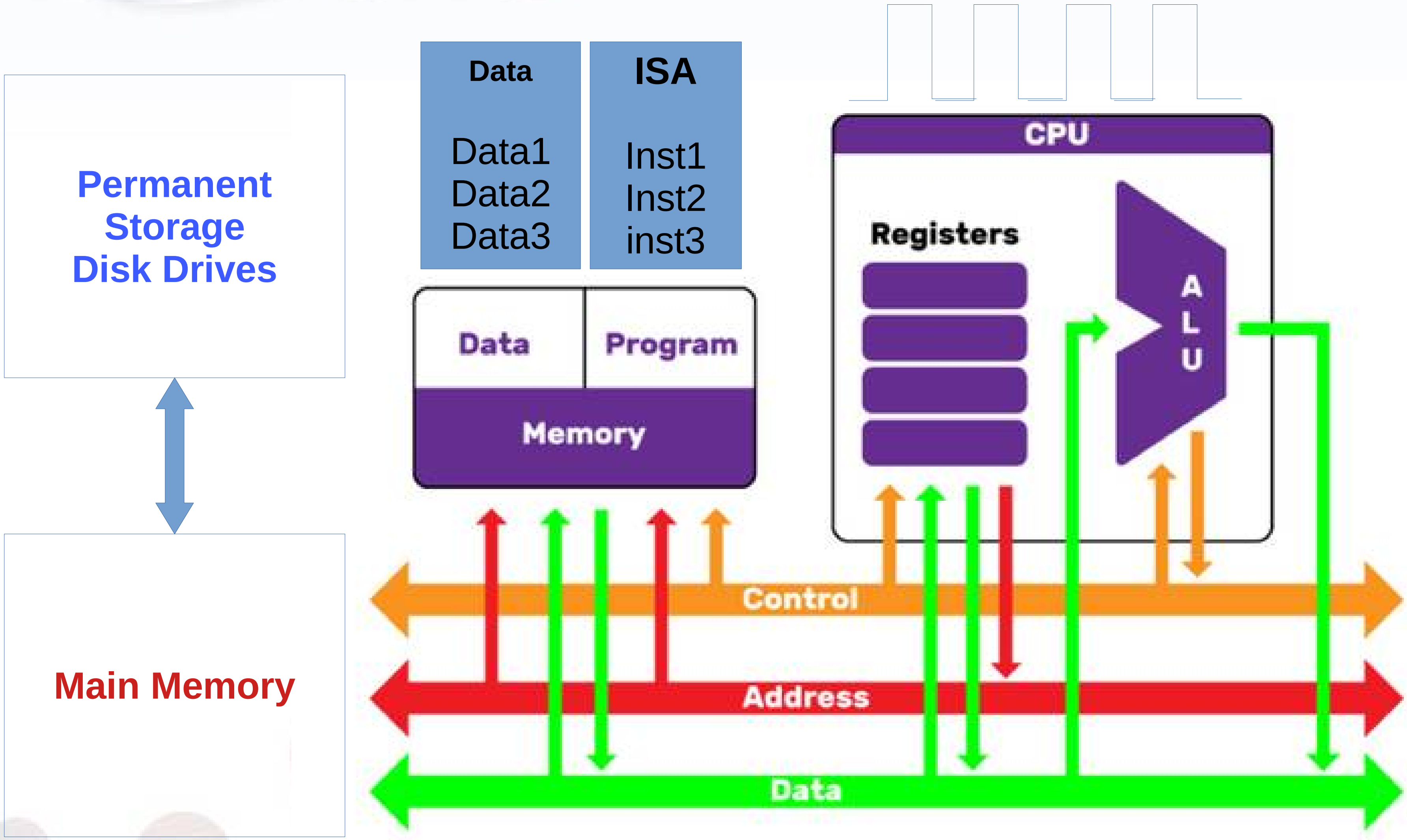
The Evolution of Computing: From Hardware Foundations to Supercomputing Power

- **Microprocessors (Proprietary Era):** Revolutionized automation and digital control, laying the foundation for modern computing.
- GCC (GNU Compiler Collection): Transformed software development by introducing open-source compilers.
- Linux: Made open-source operating systems mainstream, reshaping the computing industry.
- **Open-Source Internet & Networking Protocols (TCP/IP, DNS, HTTP, 5G): Enabled global digital connectivity.**
- **Mathematical Models, Development Frameworks, and Open Datasets: Are revolutionizing AI and computational intellectuality.**
- **Cloud Computing: Democratized access to computing resources, making them available on demand.**
- **Supercomputing: Pushed the boundaries of scalable performance, powering advanced research and innovation.**

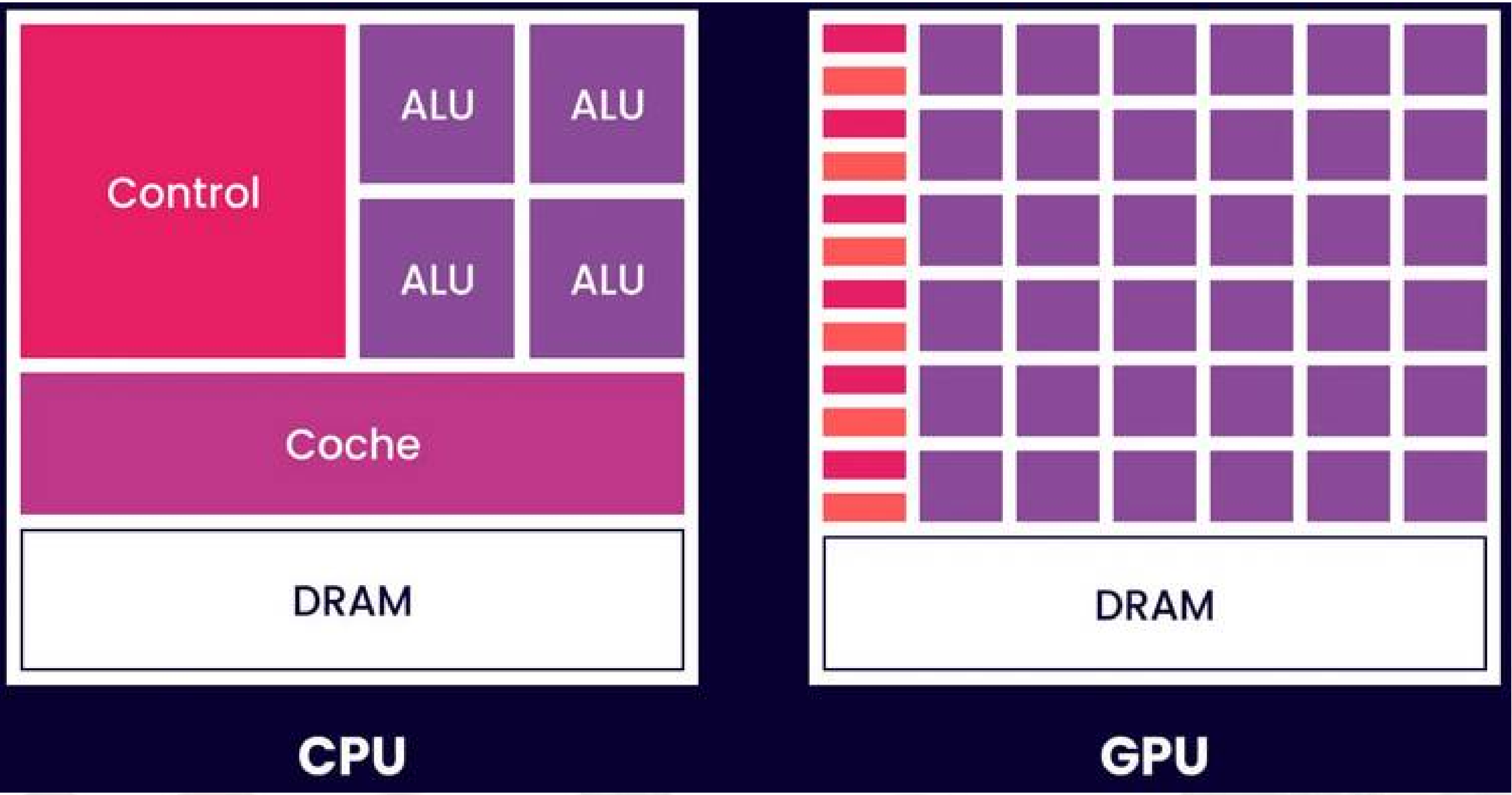
High Performance Computing

- Basics
- **Hardware Stack**
- Software Stack
- Programming Model

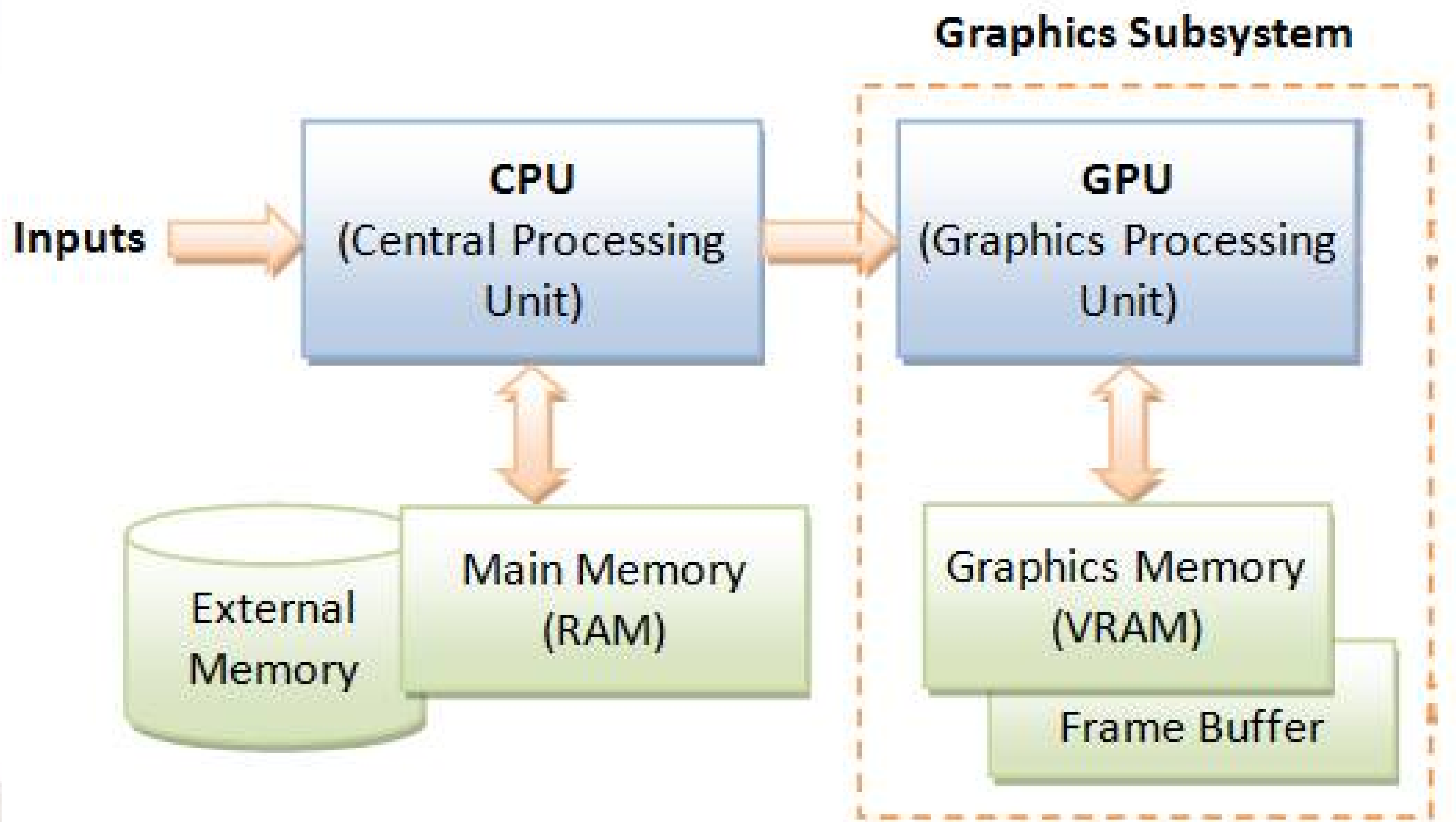
Basics of Data Processing



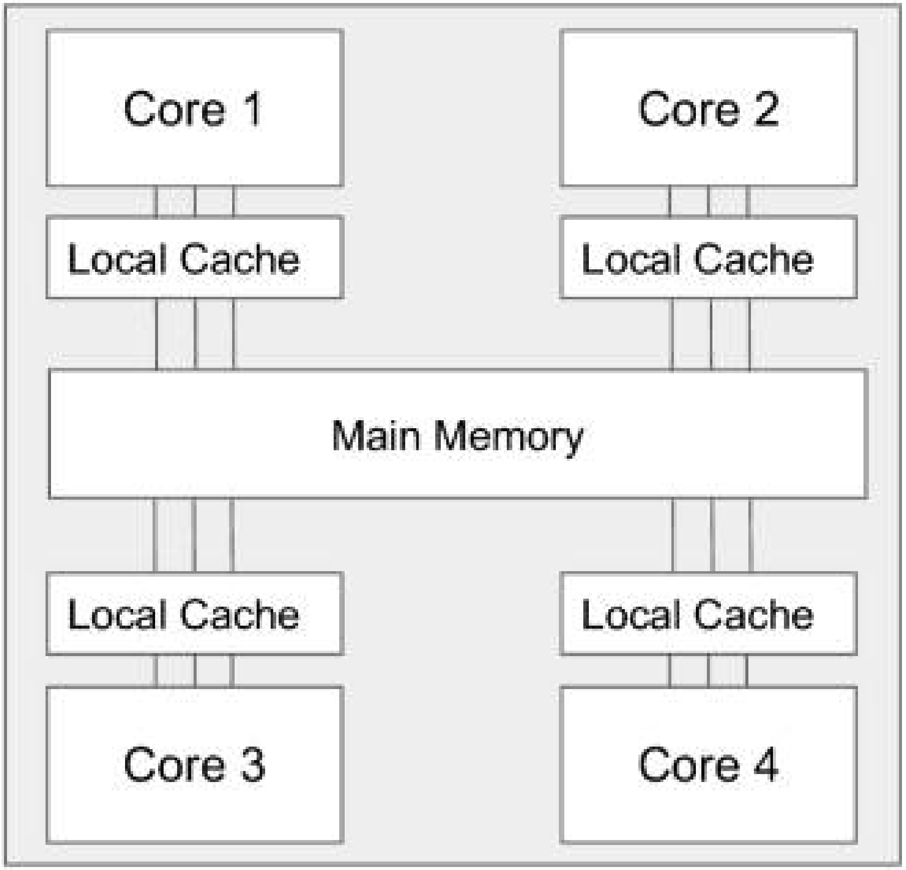
High Performance Computing



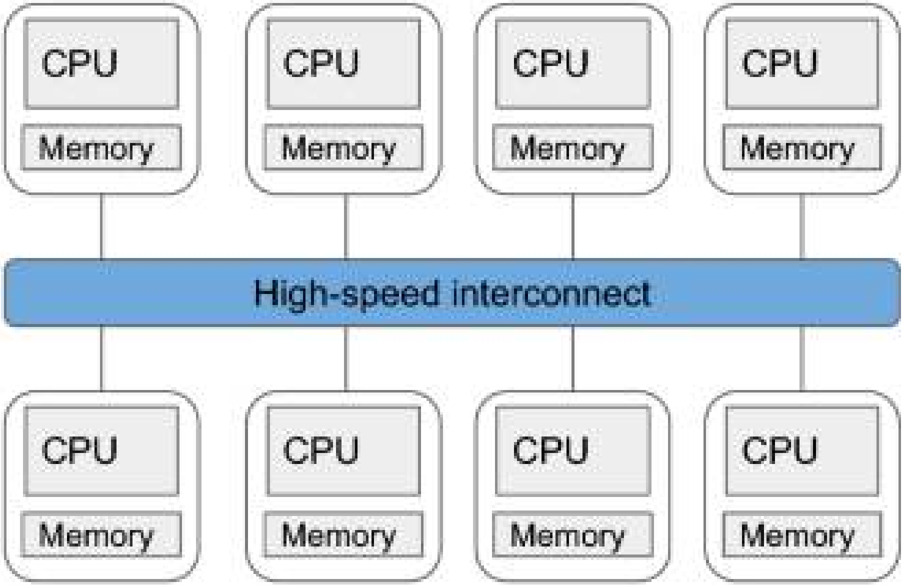
Offloading Acceleration



High Performance Computing



(a)

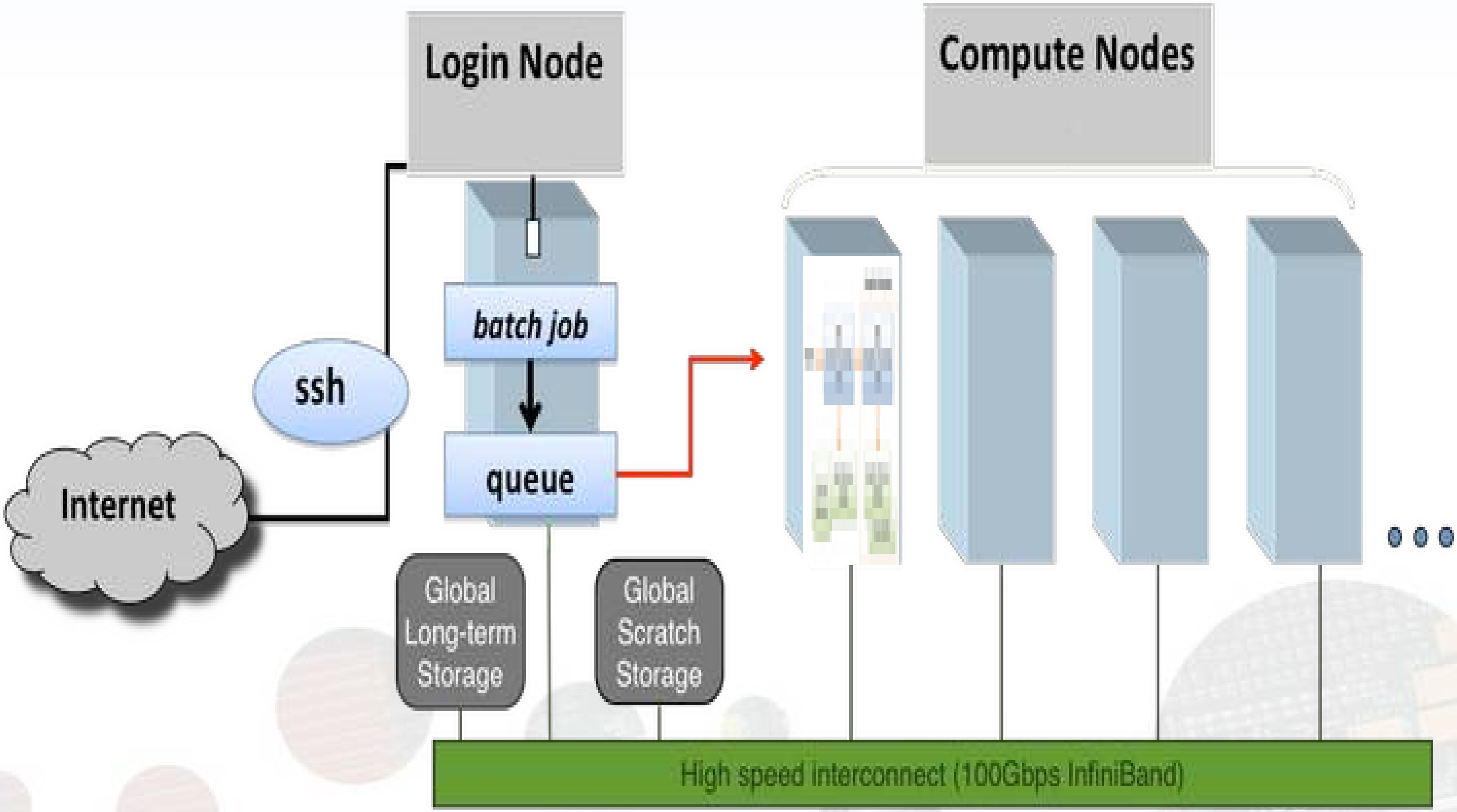


(b)

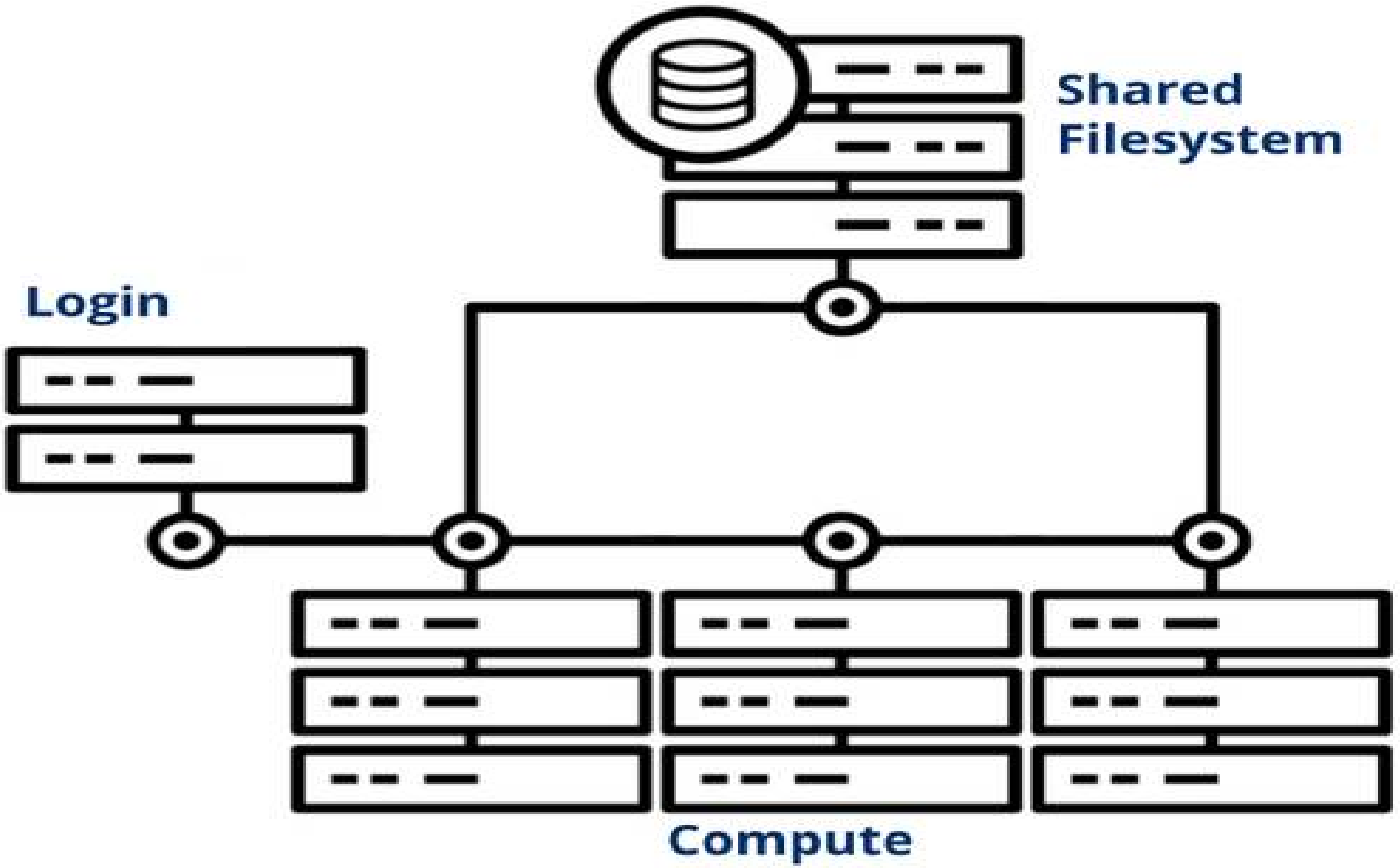
Core Components of HPC

- **Compute Nodes**
  - Individual servers machines designed to perform calculations.
  - Interconnected through a cluster that allows them to work in sync on larger problems.
- **Storage Systems**
  - Robust storage solution to handle the massive data generated and processed.
  - High-speed storage systems such as parallel file systems allow efficient access and retrieval of data.
- **Networking Infrastructure:**
  - A high-speed network for connecting the compute nodes and storage systems in an HPC cluster.
  - Technologies like InfiniBand and Ethernet play a critical role in minimizing latency and maximizing data transfer rates.

Working Example



BigData: Shared File System



High Performance Computing

- Basics
- Hardware Stack
- **Software Stack**
- Programming Model

Software Stack

1) Operating system & base

– Linux (RHEL/CentOS/Alma, Rocky, Ubuntu, SLES) with NUMA, hugepages, cpufreq governor=performance.

2) Network stack (fabric + comms)

– Fabric/NIC drivers: Mellanox/NVIDIA OFED (InfiniBand), RoCEv2, or high-speed Ethernet.

– RDMA stack: rdma-core, verbs, SR-IOV (if virtualized), PFC/ECN (for RoCE).

3) Resource & job management

– Scheduler/Resource manager: SLURM (or PBS Pro, LSF).

4) Storage & filesystems

– Parallel FS: Lustre, BeeGFS, (or Spectrum Scale), plus NFS for home.

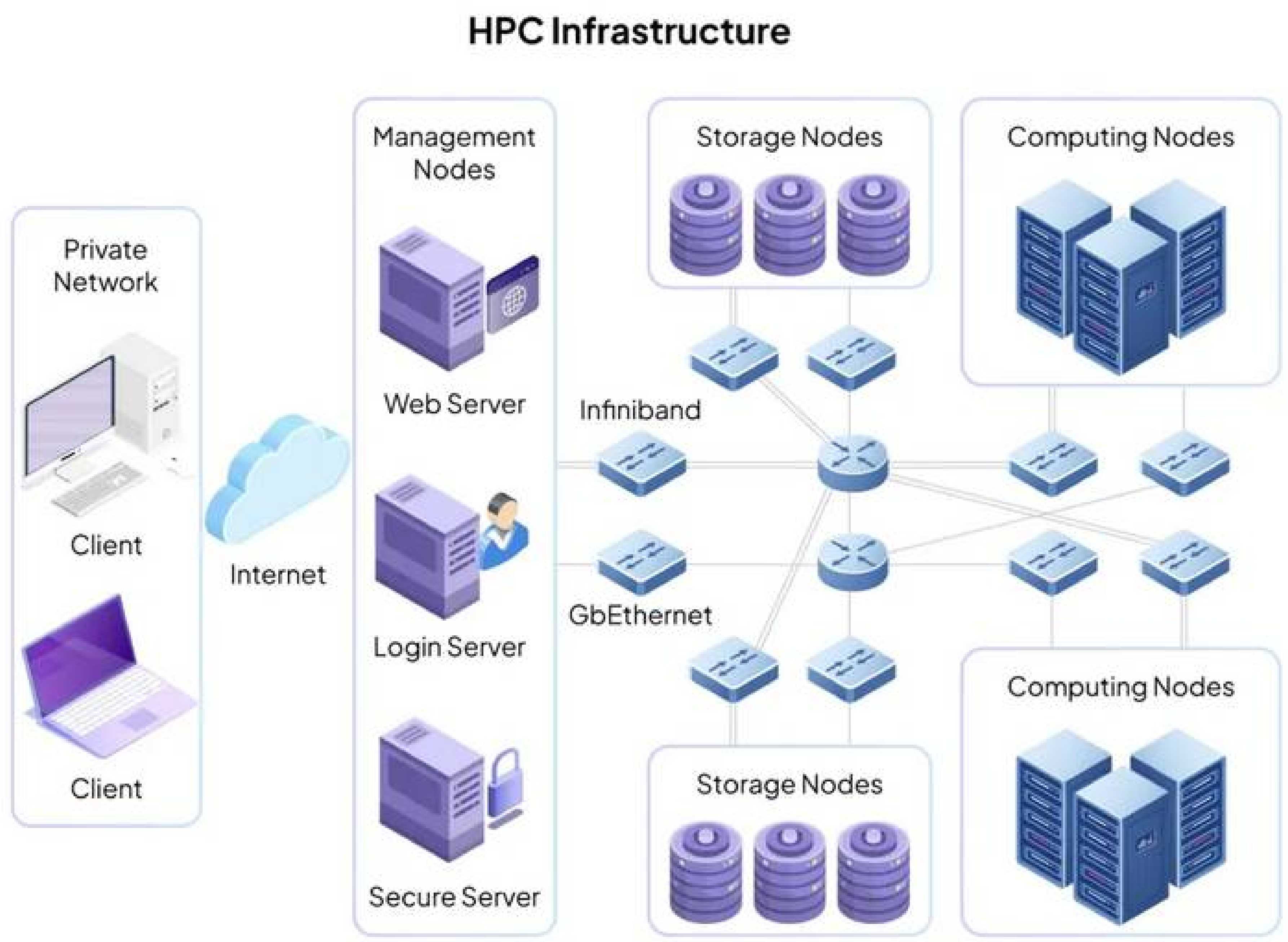
– Client tuning: read-ahead, RPC credits, MDT MDT-striping for metadata-heavy jobs.

– Burst buffers/SSD cache if available.

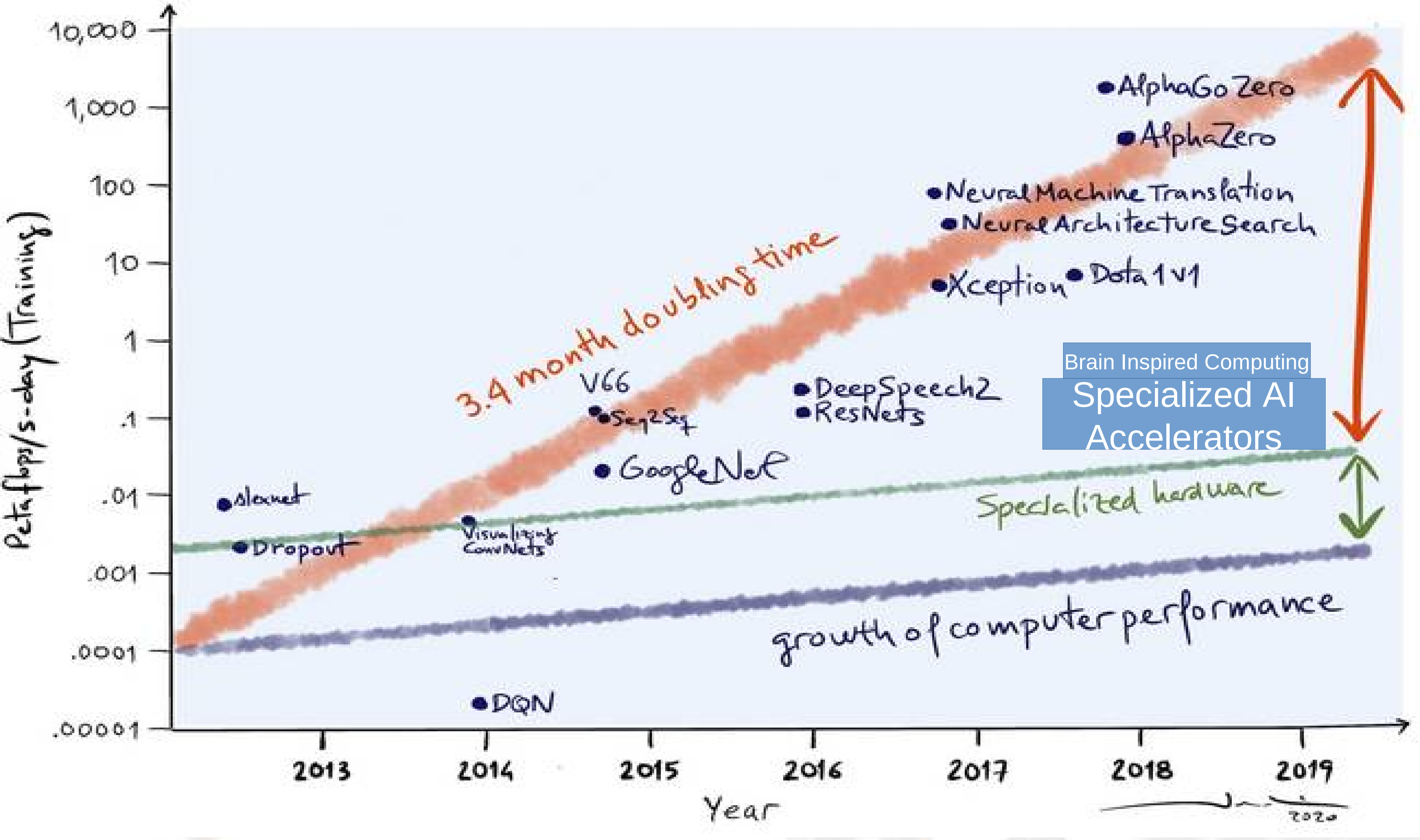
5) Parallel Programming Models

– Compilers and Libraries

– Shared and Distributed Programming Models


HPC Infrastructure
Private Network
Client
Client
Internet
Management Nodes
Web Server
Login Server
Secure Server
Storage Nodes
Infiniband
GbEthernet
Storage Nodes
Computing Nodes
Computing Nodes

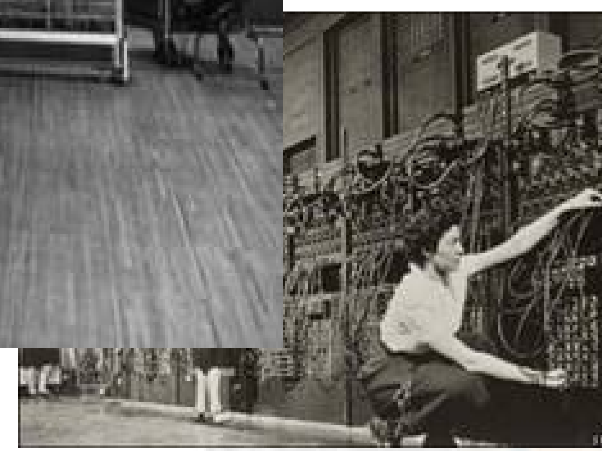
AI and Specialized Accelerators Performance Gap



High Performance Computing

- Basics
- Hardware Stack
- Software Stack
- **Programming Model**

Programming in the Early Mainframe Era



Before keyboards and screens: Early mainframes were programmed by toggling switches and monitoring lights.

Programming HPC

Mainframe Era (1940s–1960s)

- Programming via switches, punched cards, assembly language

Vector & Scientific Supercomputers (1970s–1980s)

- Programming with Fortran, vectorization

Cluster & Parallel Era (1990s–2000s)

- MPI (Message Passing Interface) and OpenMP became standard

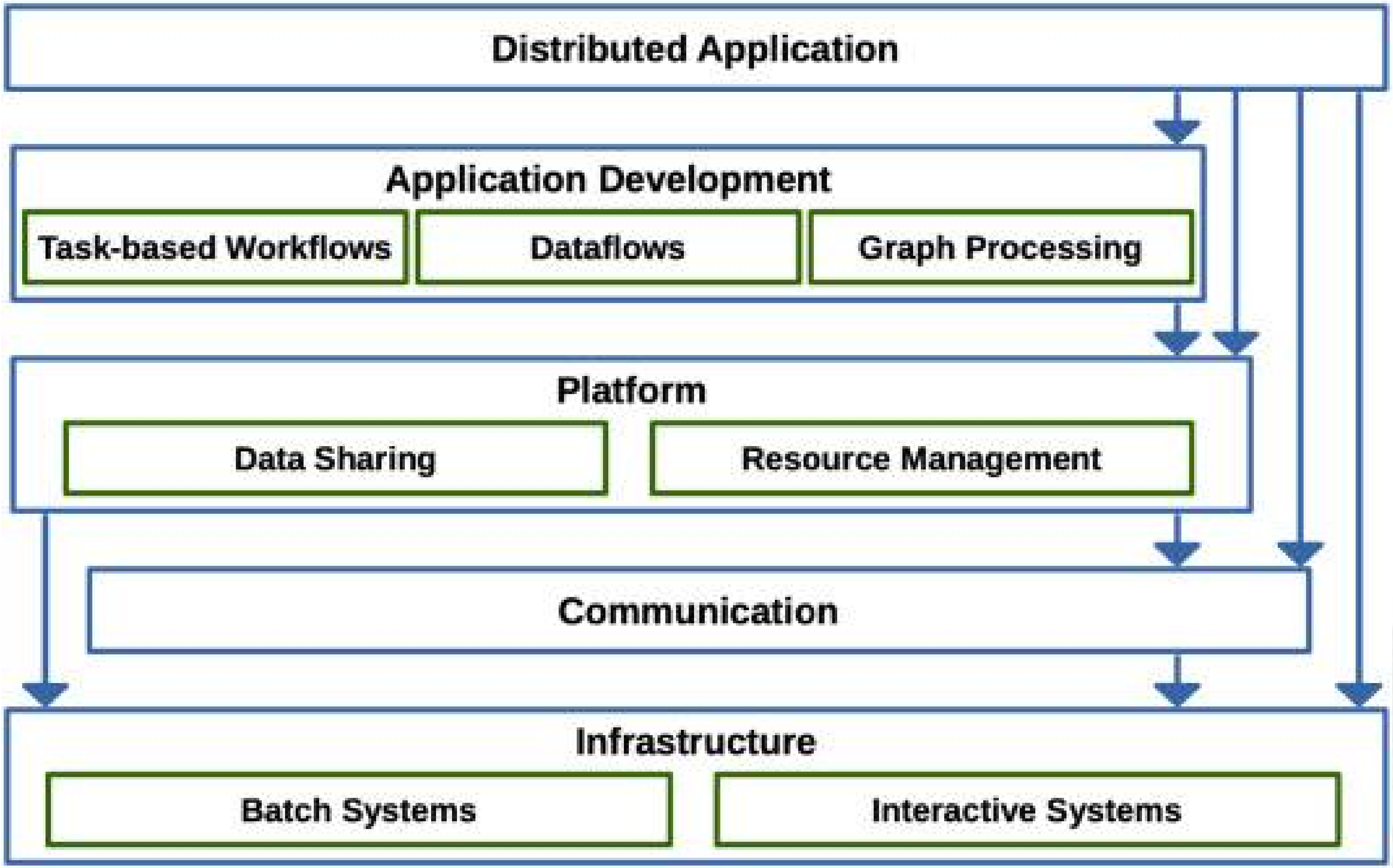
Petascale & Specialized Acceleration (2010s)

- Programming models expanded: CUDA, OpenACC, SYCL, OpenMP offload

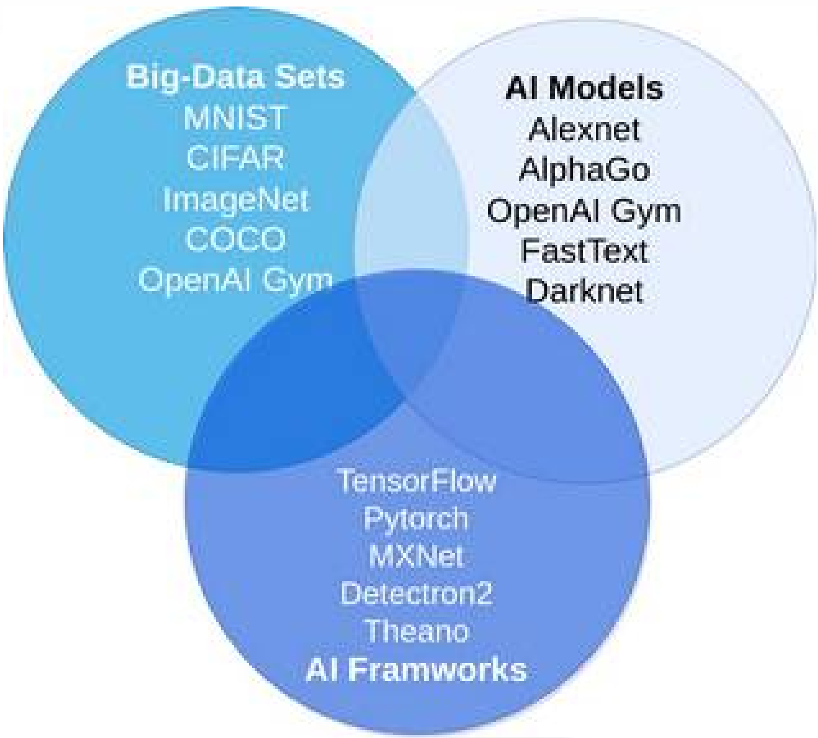
Exascale & Heterogeneous Era (2020s–present)

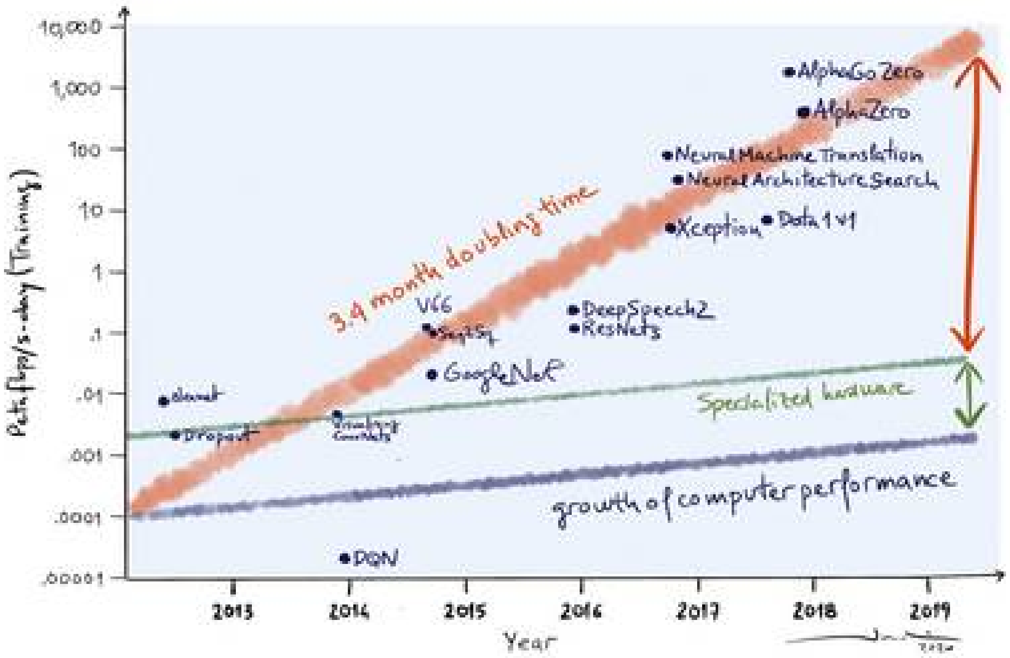
- Programming Models= managing heterogeneous nodes (CPUs, GPUs, AI accelerators)

Data Processing Stack: Distributed Programming Models



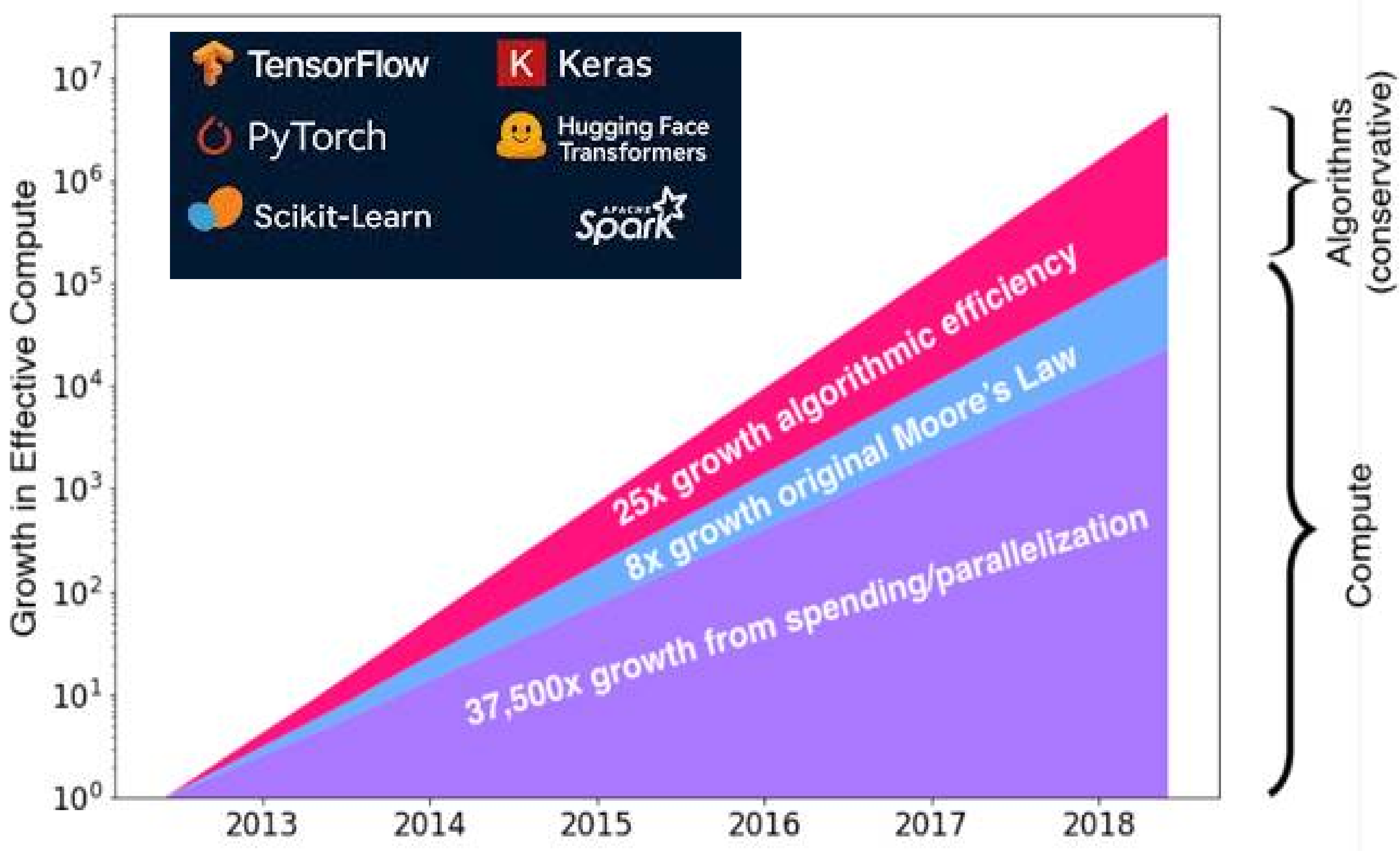
Open-Source Software:
Compilers, Mathematical Algorithms and Data-sets





Mathematical Algorithms, Big datasets, and open-source DL framework, play an important role to create "big" algorithms.

Meeting Computational Demand: Solutions in Parallel Processing



Acceleration by using Distributed and Parallel Computing

Google Multilingual translation Model 600 Billion parameters takes 22 Years to get trained on 1TPU Machine.

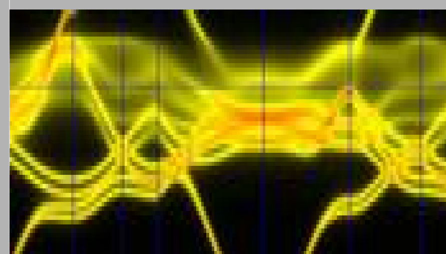
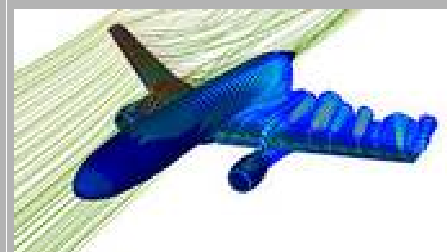
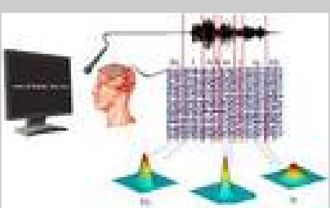
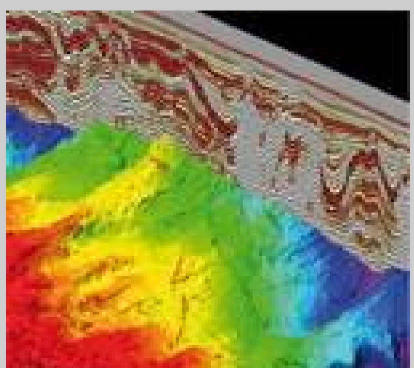
While using Google distributed the training over 2048 TPUs and achieved results in only 4 days.

Jordi: https://towardsdatascience.com/artificial-intelligence-is-a-supercomputing-problem-4b0edbc2888d

- Mankind Progress, Industrial and Academic Revolution
- Age of Big Data and AI
- Supercomputing! Revolutionizing the World
- **Namal Centre for AI and BigData: Supercomputing Facility**

HPC Applications in Pakistan

- **Representative application domains requiring more than a Desktop PC Performance (50+ Workshops and Seminars)**

Biomedical [Alpha Genomic] 	Control and Simulation 	Aerodynamics 	Mechanical Systems Modeling and Simulation 
Brain Computer Interface [Riphah NewZeland College of Chiropractic] 	Earth Sciences (QAU) 	Big-Data Analytic 	Artificial Intelligence 

Centre for AI and BigData

The mission of the Centre for AI & Big Data is to solve local challenges for complex and compute intensive data, using cutting-edge smart solutions and state-of-the-art high-performance computing technologies that drive sustainable industrial development and economic growth.

The center holds infrastructure and capabilities that are only available in Pakistan:

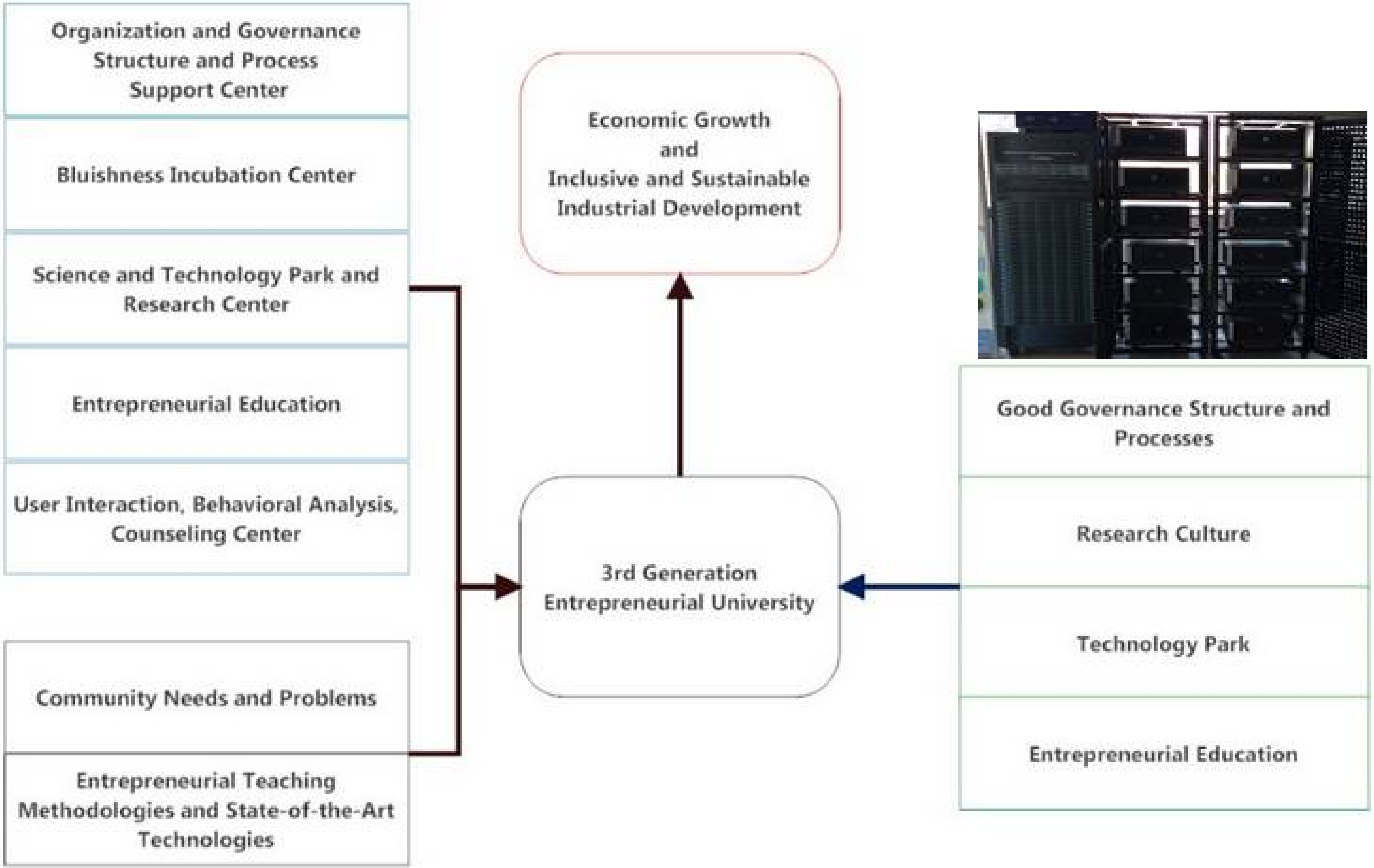
Indigenous Supercomputing Cluster capable of providing 1120 **TeraFLOPS performance.**

High-Performance Computing, Cloud Computing, and Data Center **Experts.**

Distributed Artificial Intelligence and **Parallel Programming Developers.**

Secure Processor-based Systems and computer architecture expertise.

Heterogeneous Parallel Programming Framework for **Artificial Intelligent Applications.**


Organization and Governance Structure and Process Support Center
Bluishness Incubation Center
Science and Technology Park and Research Center
Entrepreneurial Education
User Interaction, Behavioral Analysis, Counseling Center
Community Needs and Problems
Entrepreneurial Teaching Methodologies and State-of-the-Art Technologies
Economic Growth and Inclusive and Sustainable Industrial Development
3rd Generation Entrepreneurial University
Good Governance Structure and Processes
Research Culture
Technology Park
Entrepreneurial Education

Problem

Digital

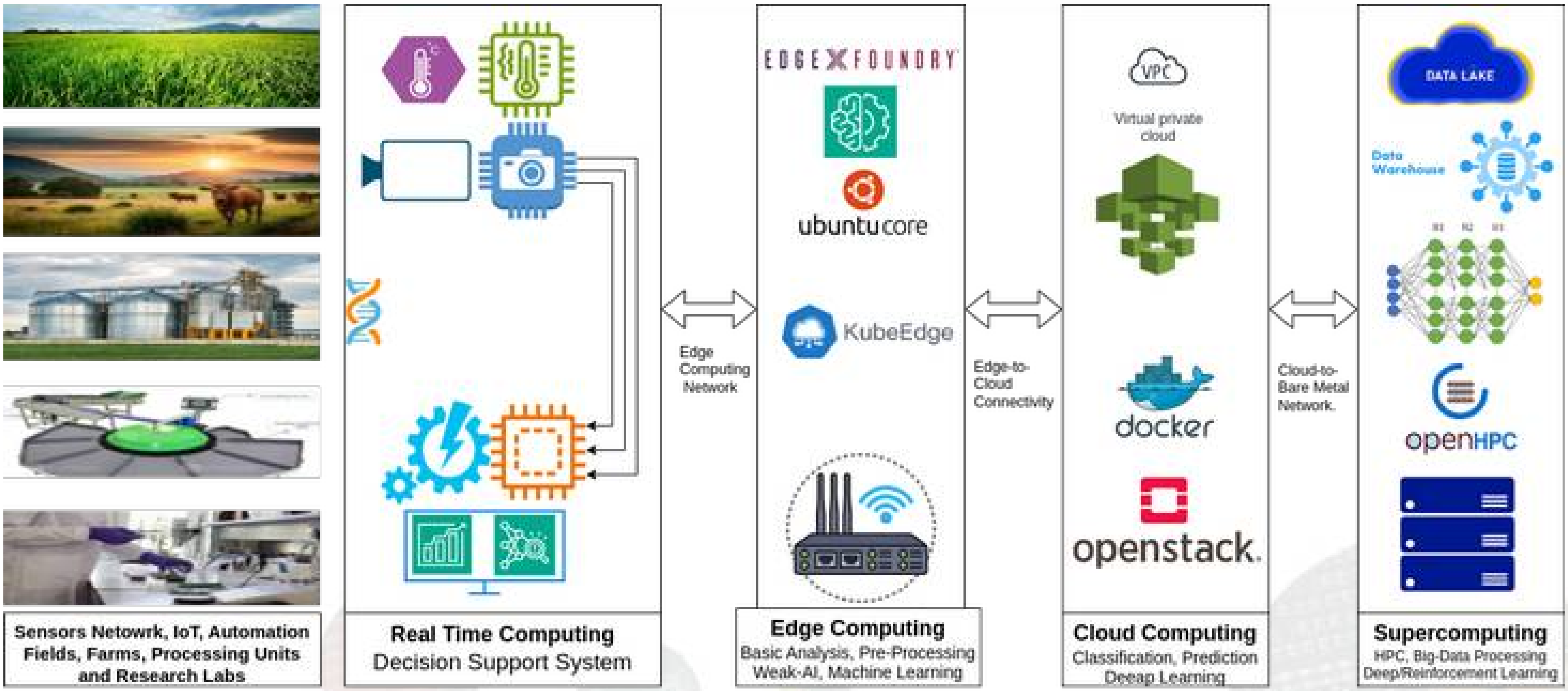
Data
Software
Hardware
Front-end

Low-Performance
Technology
High-Performance

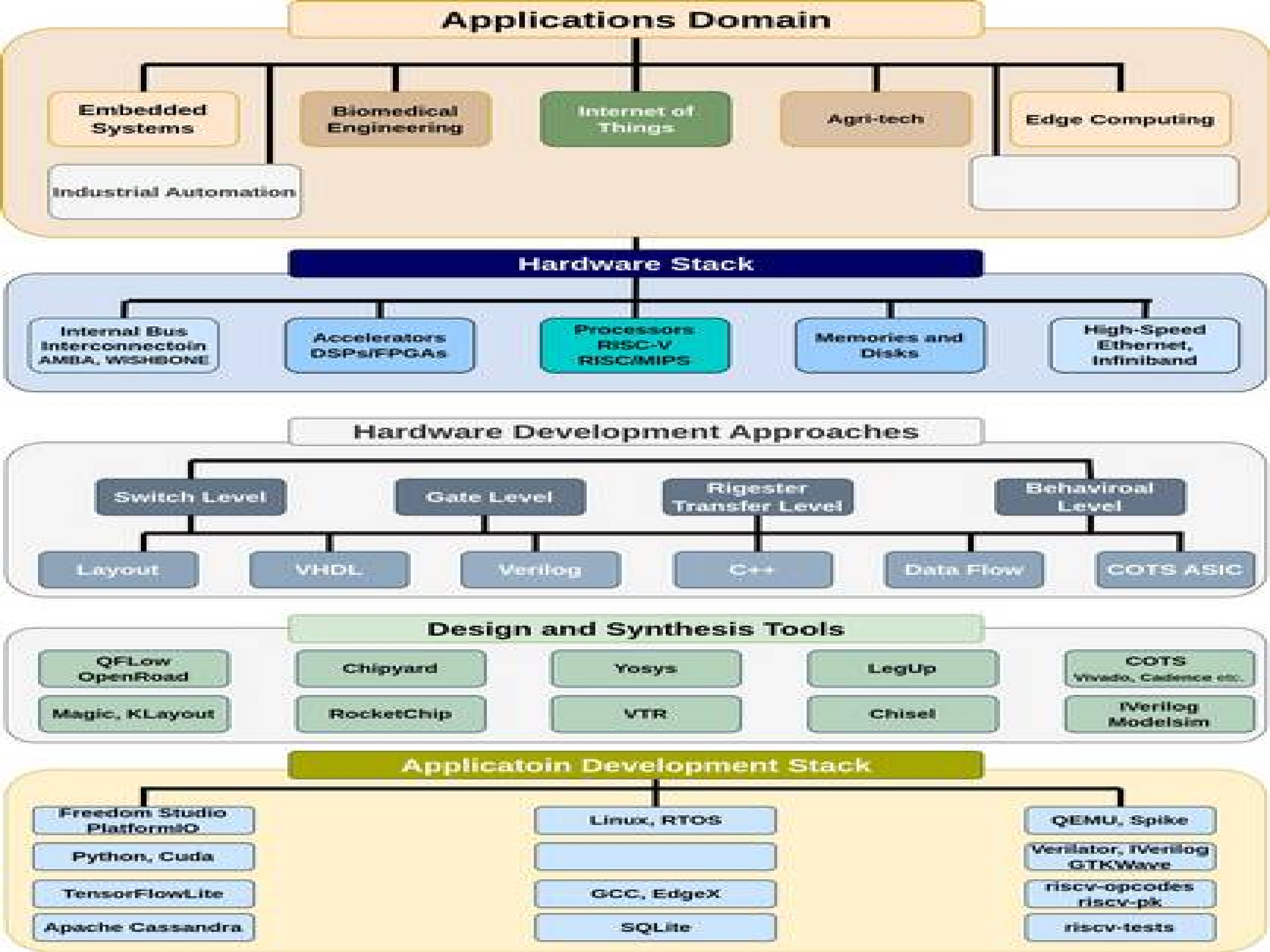
Embedded/Edge

Cloud

Bare-Metal



Baremetal: ssh namal-hpc@hpc.computingpark.com
Cloud Application: http://cloud.computingpark.com/
Data Center: https://data.computingpark.com



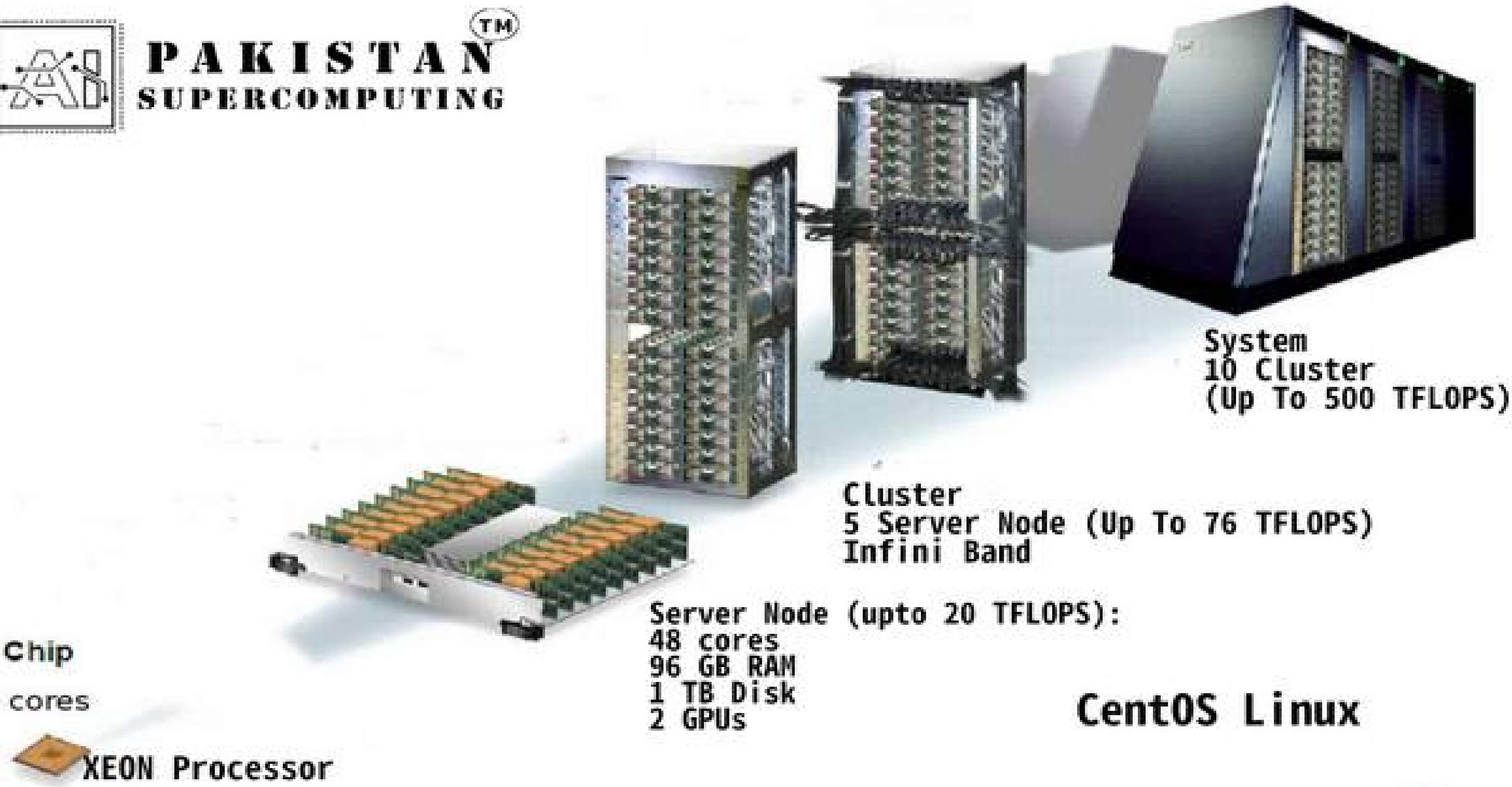
Applications Domain
Embedded Systems
Biomedical Engineering
Internet of Things
Agri-tech
Edge Computing
Industrial Automation
Hardware Stack
Internal Bus Interconnectoin AMBA, WISHBONE
Accelerators DSPs/FPGAs
Processors RISC-V RISC/MIPS
Memories and Disks
High-Speed Ethernet, Infiniband
Hardware Development Approaches
Switch Level
Gate Level
Rigester Transfer Level
Behaviroal Level
Layout
VHDL
Verilog
C++
Data Flow
COTS ASIC
Design and Synthesis Tools
QFLow OpenRoad
Chipyard
Yosys
LegUp
COTS Vivado, Cadence etc.
Magic, KLayout
RocketChip
VTR
Chisel
IVerilog Modelsim
Applicatoin Development Stack
Freedom Studio PlatformIO
Python, Cuda
TensorFlowLite
Apache Cassandra
Linux, RTOS
GCC, EdgeX
SQLite
QEMU, Spike
Verilator, IVerilog GTKWave
riscv-opcodes riscv-pk
riscv-tests

AI and BigData Applications

- **Bare-Metal and Containerized Cluster Infrastructure:**
  - Distributed Hardware Interfacing, Network Configuration and Distributed Computing Software Deployment
- **Data Center and Cloud Infrastructure:**
  - Storage systems, networking equipment, and software configuration
- **AI Applications for Scientific and Engineering Problems**
  - Distributed AI applications for multi-node bare-metal system
- **HPC Application Parallel Programming**
  - Heterogeneous multi-node parallel processing using parallel programming models

Developing Supercomputing for AI

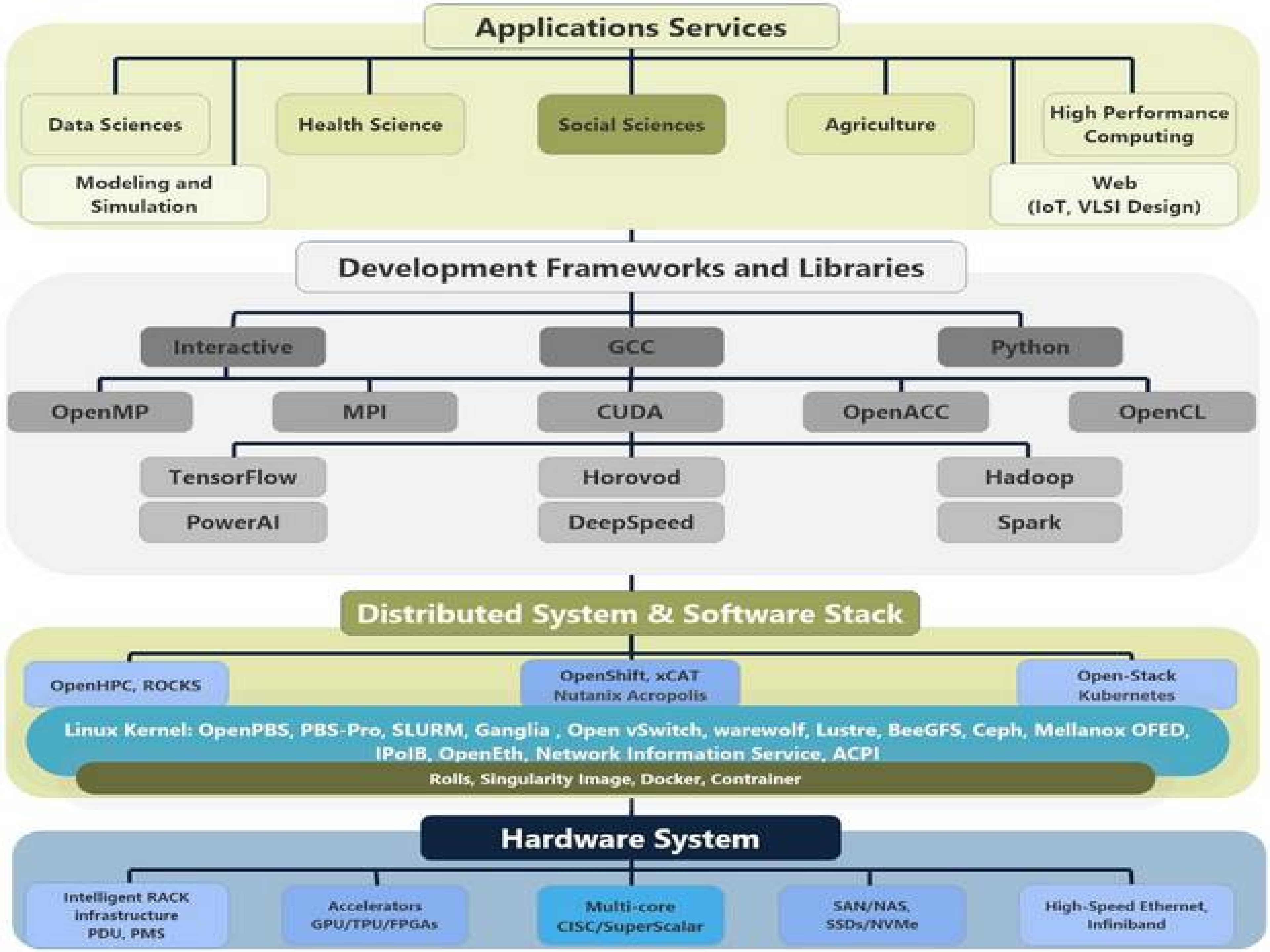
PAKISTAN SUPERCOMPUTING ™



CentOS Linux

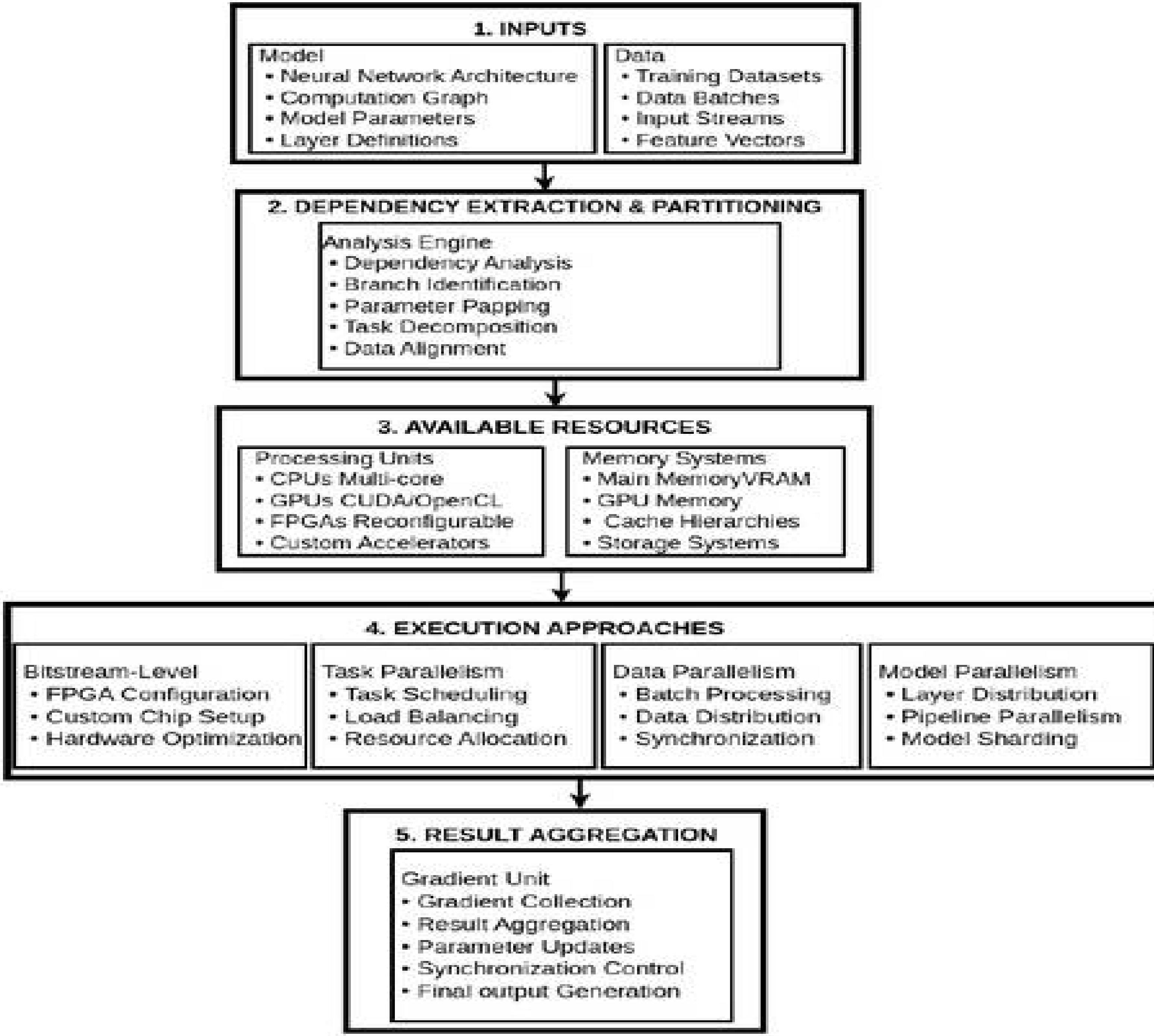
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

HEC


Applications Services
Data Sciences
Health Science
Social Sciences
Agriculture
High Performance Computing
Modeling and Simulation
Web (IoT, VLSI Design)
Development Frameworks and Libraries
Interactive
GCC
Python
OpenMP
MPI
CUDA
OpenACC
OpenCL
TensorFlow
Horovod
Hadoop
PowerAI
DeepSpeed
Spark
Distributed System & Software Stack
OpenHPC, ROCKS
OpenShift, xCAT Nutanix Acropolis
Open-Stack Kubernetes
Linux Kernel: OpenPBS, PBS-Pro, SLURM, Ganglia , Open vSwitch, warewolf, Lustre, BeeGFS, Ceph, Mellanox OFED, IPoIB, OpenEth, Network Information Service, ACPI
Rolls, Singularity Image, Docker, Contrainer
Hardware System
Intelligent RACK infrastructure PDU, PMS
Accelerators GPU/TPU/FPGAs
Multi-core CISC/SuperScalar
SAN/NAS, SSDs/NVMe
High-Speed Ethernet, Infiniband

		HPC	AI	BigData
Apps...	Boundary Interaction Services	- **Remote access**: X2Go, Apache Guacamole, OpenVNC - **Visualization**: ParaView, VisIt (via X11 forwarding or browser) - **Secure access**: OpenSSL, SSH with key-based login - **User portals**: Cockpit, Slurm-Web	- **Model serving**: FastAPI, Gradio, Streamlit (self-hosted) - **Web access**: JupyterHub, VS Code Server - **Authentication**: OpenSSL/TLS, OAuth2 via open-source IdPs (e.g., Keycloak)	- **Dashboards**: Grafana, Superset, Apache ECharts - **Remote data access**: WebHDFS, OpenNMS - **Visualization tools**: Apache Zeppelin, D3.js, Plotly Dash - **Secure access**: OpenSSL, SSH
Middleware & MGMT	Processing Services	**Domain Specific frameworks** [e.g. PETSc], **Batch processing of large tightly coordinated parallel jobs** [100s - 10000s of processes communicating frequently with each other]	**DNN training & interference frameworks** [e.g. Caffe, Tensorflow, Theano, Neon, Torch], **DNN numerical libraries** [e.g. dense LA]	**Machine Learning** (traditional) [e.g. Mahout, Scikit-learn, BigDL], **Analytics/ Statistics** [e.g. Python, ROOT, R, Matlab, SAS, SPSS, Sci-Py], **Iterative** [e.g. Apache Hama], **Interactive** [e.g. Dremel, Drill, Tez, Impala, Shark, Presto, BlinkDB, Spark], **Batch / Map Reduce** [e.g. MapReduce, YARN, Sqoop, Spark], **Real-time / streaming** [e.g. Flink, YARN, Druid, Pinot, Storm, Samza, Spark]
	Model / Infromation Management Services	**Data Storage**: Parallel File Systems [e.g. Lustre, GPFS, BeeGFS, PanFS, PVFS], **I/O libraries** [e.g. HDF5, PnetCDF, ADIOS]	**Data Storage** [e.g. HDFS, Hbase, Amazon S3, GlusterFS, Cassandra, MongoBD, Hana, Vora]	**Serialization** [e.g. Avro], **Meta data** [e.g. HCatalog], **Data Ingestion & Integration** [e.g. Flume, Sqoop, Apache Nifi, Elastic Logstash, Kafka, Talend, Pentaho], **Data Storage** [e.g. HDFS, Hbase, Amazon S3, GlusterFS, Cassandra, MongoBD, Hana, Vora], **Cluster Mgmt** [e.g. YARN, MESO]
	Communication Services	**Messaging & Coordination** [e.g. MPI/PGAS, direct fabric access], **Threading** [e.g. OpenMP, task-based models]	**Messaging & Coordination** [e.g. Machine Learning Scaling library(MLSL)]	**Messaging** [e.g. Apache Kafka (streaming)]
	Workflow / Task Services	**Conventional compiled languages** [e.g. C/C++/Fortran], **Scripting languages** [e.g. Python, Julia]	**Scripting languages** [e.g. Python]	**Workflow & Scheduling** [e.g. Oozie], **Scripting languages** [e.g. Leras, Mocha, Pig, JAQL, Python, Java, Scala
System SW	System Management & Security Services	**Domain numerical libraries** [e.g. PETSc, ScaLAPACK, BLAS, FFTW,...], **Performance & debugging** [e.g. DDT, Vampire], **Accelerator APIs** [e.g. CUDA, OpenCL, OpenACC] **Data Protection** [e.g. System SSS, OS/PFS file access control] **Batch scheduling** [e.g. SLURM], **Cluster management** [e.g. OpenHPC], **Container Virtualization** [e.g. Docker], **Operating System** [e.g. Linux OS Variant]	**Batching for training** [built into DL frameworks], **Reduced precision** [e.g. interference engines], **Load distribution layer** [e.g. Round robin/load balancing for interference], **Accelerator APIs** [e.g. CUDA, OpenCL], **Hardware Optimization Libraries** [e.g. cuDNN, MKL-DNN, etc.] **Virtualisation** [e.g Dockers, Kubernetes, VMware, Xen, KVM, HyperX], **Operating System** [e.g. Linux (RedHat, Ubuntu, etc.), Windows]	**Distributed Coordination** [e.g. ZooKeeper, Chubby, Paxos], **Provisioning, Managing & Monitoring** [e.g. Ambari, Whirr, BigTop, Chukwa], **SVM systems** [e.g. Google Sofia, libSVM, svm-py,...], **Hardware Optimization Libraries** [e.g. DAAL, DPDK, MKL, etc.] **Virtualization** [e.g. Dockers, Kubernetes, VMware, Xen, KVM, HyperX], **Operating System**[e.g. Linux (RedHat, Ubuntu, etc.), Windows]
Hardware	Infrastructure	**Local storage** [e.g. Storage & I/O nodes, NAS]; **Servers**[e.g. CPU & Memory [Gen Purpose CPU nodes, GPUs, FPGAs]; **Network** [e.g. Infiniband & OPA fabrics]	**Local storage** [e.g. Local storage or NAS/SAN]; **Services** [e.g. CPU & Memory [Gen Purpose CPU + GPU/FPGA, TPU]; **Network** [e.g. Ethernet]	**Local Storage** [e.g. Direct attached Storage]; **Services** [e.g. CPU & Memory, [Gen Purpose CPU hyper-convergent nodes]; **Network** [e.g. Ethernet fabrics]

High Performance Computing	Cloud Computing	Big Data, Edge Computing
Access & Security		
X2Go, Guacamole, OpenVNC, SSH, Cockpit, Slurm-Web	FastAPI, Gradio, Triton, JupyterHub, Keycloak (OAuth2), VS Code Server	Grafana, Superset, Zeppelin, WebHDFS, SSH + Keycloak
Frameworks and Libraries		
PETSc, OpenHPC, Spack, SLURM, OpenMPI	PyTorch, TensorFlow, ONNX, JAX, MLFlow, HuggingFace, Kubeflow	Spark, Flink, Mahout, MLlib, H2O.ai, Scikit-learn, Druid
Storage and Messaging		
Lustre, BeeGFS, GlusterFS, CephFS, HDF5, ADIOS2, NetCDF	HDFS, MinIO, MongoDB, Redis, Ceph, DVC, Weights & Biases	HDFS, Iceberg, Delta Lake, MinIO, Kafka, Pulsar, Zookeeper
Programming & Languages		
C, C++, Fortran, Python, Julia, OpenMP, Spack	Python, Julia, R, YAML (Kube), TensorBoard, Torch Profiler	Java, Scala, Python, Pig, HiveQL, Airflow DAGs
DevOps, MLOps & CI/CD		
Spack, EasyBuild, Singularity, Ansible, Warewulf, Podman	Kubeflow, MLFlow, Metaflow, DVC, Docker, K3s, KServe	Jenkins, ArgoCD, GitLab CI, NiFi, StreamSets, Airflow
Containerization & Orchestration		
Podman, Singularity, Apptainer, KVM, SLURM	Docker, Kubernetes, K3s, Kubeflow Pipelines	Docker, Kubernetes, Helm, MicroK8s, Oozie
OS, Virtualization, Networking		
Rocky Linux 9.x, AlmaLinux, KVM, Infiniband, RoCEv2	Ubuntu 22.04+, Debian 12+, KVM, K3s, RDMA, Ethernet	Ubuntu, Debian, Kubernetes CNI (Calico, Flannel), Ethernet
Monitoring & Performance		
Prometheus, DCGM, Ganglia, Grafana, Valgrind, Vampir, DDT	Prometheus, Grafana, TensorBoard, PyTorch Profiler	Prometheus, Grafana, Alertmanager, Kafka Monitor
Hardware & Services		
CPU, GPU, FPGA, NAS/SAN, NVMe, High-speed fabrics	CPU, GPU, TPU, Edge Devices (Jetson), NVMe, Ceph	CPU, GPU, Hyper-converged nodes, MinIO, Ceph, HDFS


1. INPUTS
Model
• Neural Network Architecture
• Computation Graph
• Model Parameters
• Layer Definitions
Data
• Training Datasets
• Data Batches
• Input Streams
• Feature Vectors
2. DEPENDENCY EXTRACTION & PARTITIONING
Analysis Engine
• Dependency Analysis
• Branch Identification
• Parameter Papping
• Task Decomposition
• Data Alignment
3. AVAILABLE RESOURCES
Processing Units
• CPUs Multi-core
• GPUs CUDA/OpenCL
• FPGAs Reconfigurable
• Custom Accelerators
Memory Systems
• Main Memory/VRAM
• GPU Memory
• Cache Hierarchies
• Storage Systems
4. EXECUTION APPROACHES
Bitstream-Level
• FPGA Configuration
• Custom Chip Setup
• Hardware Optimization
Task Parallelism
• Task Scheduling
• Load Balancing
• Resource Allocation
Data Parallelism
• Batch Processing
• Data Distribution
• Synchronization
Model Parallelism
• Layer Distribution
• Pipeline Parallelism
• Model Sharding
5. RESULT AGGREGATION
Gradient Unit
• Gradient Collection
• Result Aggregation
• Parameter Updates
• Synchronization Control
• Final output Generation

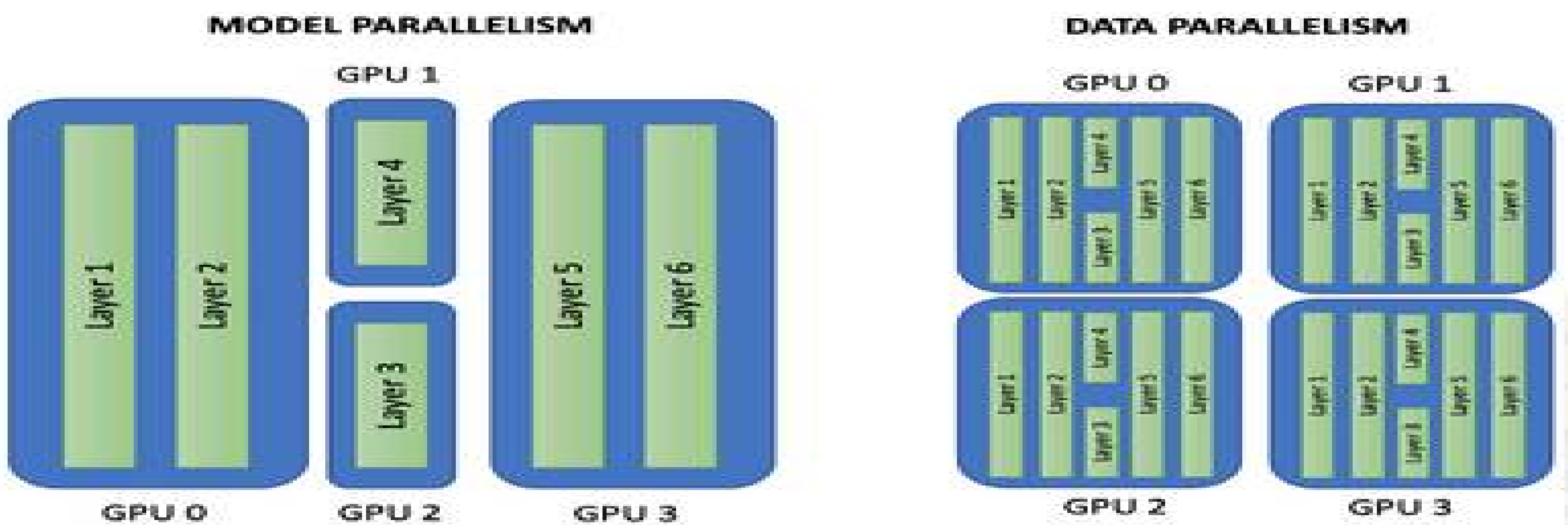
Parallel Processing

- **Model Parallelism**

  Different layers of the network distributed across different devices

- **Data Parallelism**

  Same model in every one of the GPUs, each processing a separate piece of the data, a separate portion of the mini-batch.



Visit us: ssh username@hpc.pakistansupercomputing.com

Pakistan's Number One High Performance Computing Facility:
The hardware architecture includes: 20 Nodes, 1600 Processor Core, 5 Tera Byte Main Memory, 40 TeraByte SSD, 10 Gigabit Fast Ethernet, Low Latency Switch, 40 4070TI GPU for Distributed Acceleration.
The Supercomputer is build on Rocky Linux 9.4 and features an advanced software stack including RoCE-enabled networking, Lustre parallel file system, Slurm workload manager, distributed AI and parallel programming models, with Grafana and Prometheus for real-time monitoring, and Ansible for automated deployment and management.

On-Going Projects of Relevance

1. **Supercomputing for Artificial Intelligence Applications**
   Partner: Barcelona Supercomputing Center Spain
   Pakistan Supercomputing Center
2. **Virtual Reality Platform for Rehabilitation**
   Partner: New Zealand Chiropractic Center
   Riphah Rehab Center
3. **Real-time Cattle Breed Identification System using Image Features**
   Partner: Ministry of Livestock
4. **Smart Rice Sorting Machine System**
   Partner: Alkaram Rice Engineering
5. **Secure Processor-based chip design**
   Partner: AQL Technologies
   PakASIC
6. **Smart Motor Controller**
   Partner: Khursheed Fan Gujrat
   PakASIC
7. **Footweight Analytics**
   FootAnalytic
8. **Soil Analytics**

Role and Importance of Supercompuing in Entrepreneurial Universities

by: Tassadaq Hussain

Director Centre for AI and BigData

Professor Department of Electrical Engineering

Namal University Mianwali

Collaborations:

Barcelona Supercomputing Center, Spain

European Network on High Performance and Embedded Architecture and Compilation

Pakistan Supercomputing Center