We are going through very dark days as a country. On February 6, two earthquakes of 7.7 and 7.6 magnitudes shook southeast Turkey. Unfortunately, the death toll is over 35,400 and the number of people affected is over 13 million.

My thoughts are with people who have lost loved ones, and with children who have lost their parents. What I can do right now is to focus on how we can prepare a better future for these people with the help of HPC.

KOÇ ÜNİVERSİTESİ

PAR CORE LAB

Supercomputing Software for Moore and Beyond

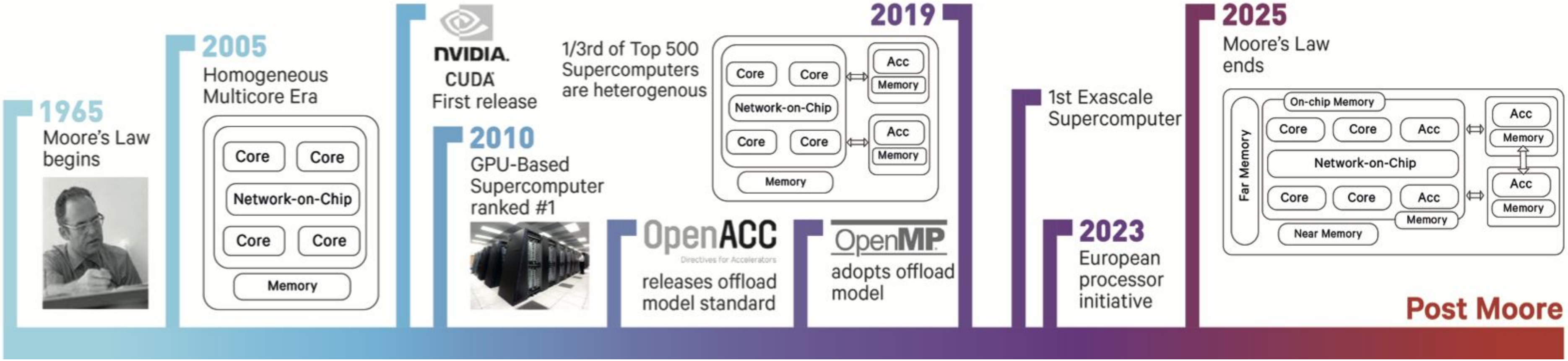
Assoc. Prof. Didem Unat
Parallel and Multicore Computing Laboratory
Koç University, Turkey

SIAG/SC • 15 Feb 2023

A brief history of Moore’s Law

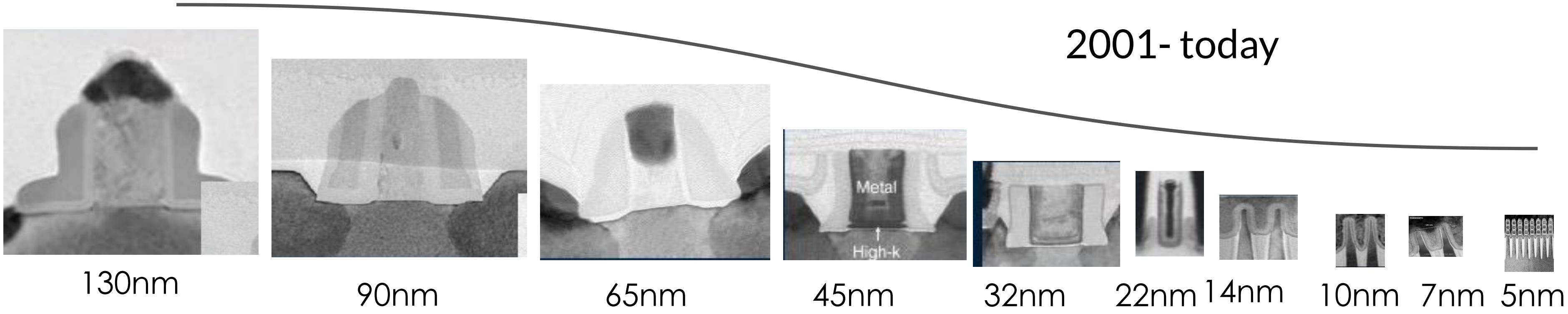
PAR CORE LAB

A brief history of Moore's Law



Lecture by Turing Award recipient Jack Dongarra

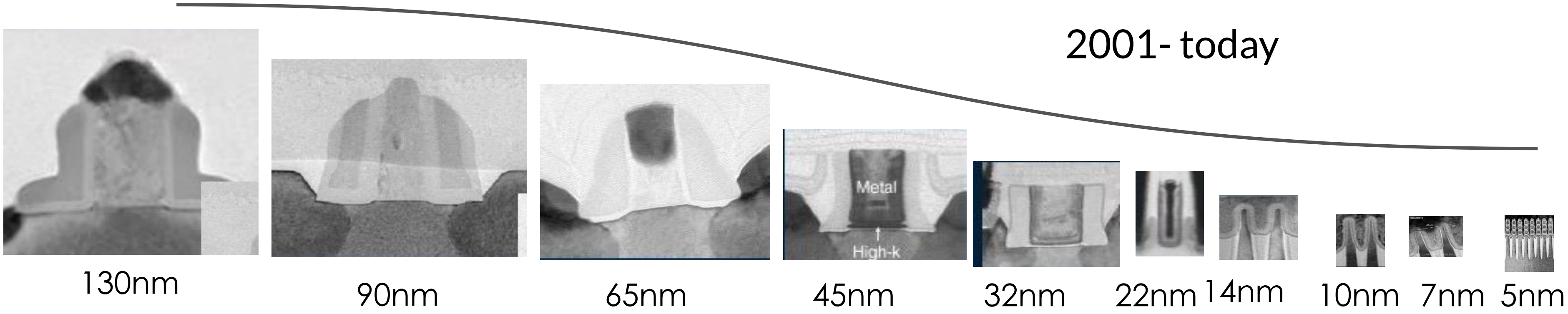
Moore’s Law



Moore's law is the observation that the number of transistors in an integrated circuit doubles about every 2 years.

[1]: Marc Horowitz, Computing’s Energy Problem (and what we can do about it), ISSC 2014, plenary
[2]: Moore: Landauer Limit Demonstrated, IEEE Spectrum 2012

Moore’s Law



- The speed of light, the atomic nature of materials and growing costs mark the end of Moore’s Law.
- Cost of manufacturing [3]
  - $170M for a 10 nm chip, $300M for a 7 nm chip, $500M for a 5 nm chip
  - For specialized chips, even higher

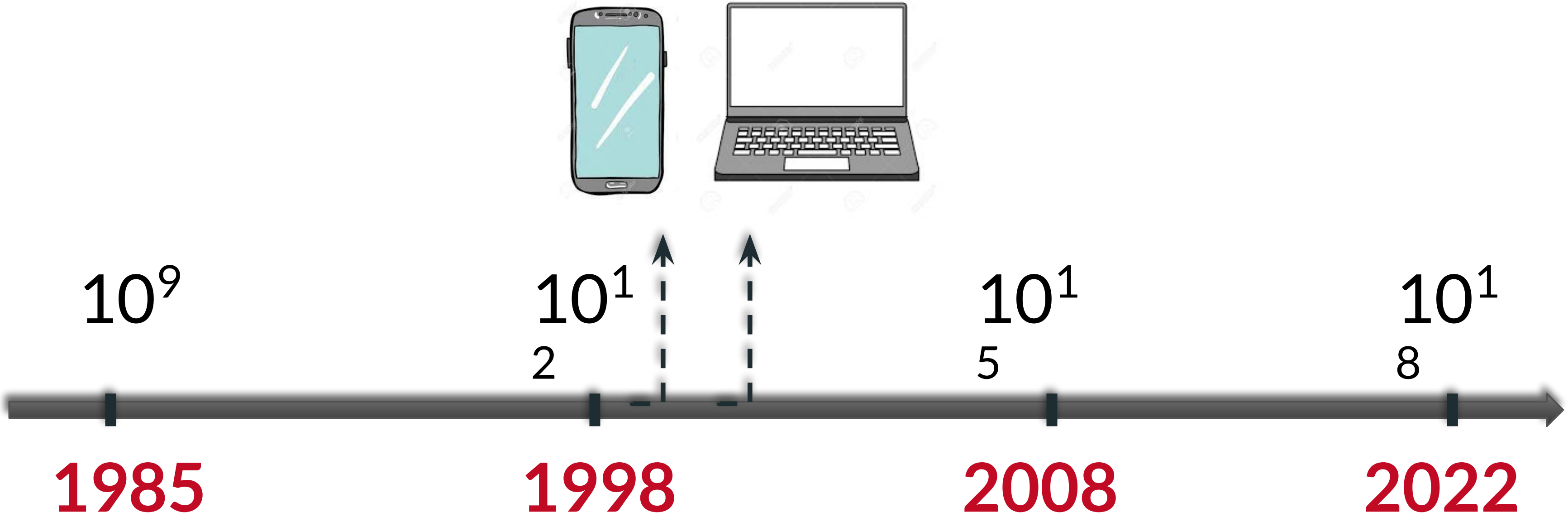
[3] https://builtin.com/hardware/moores-law

Thanks to Moore's Law

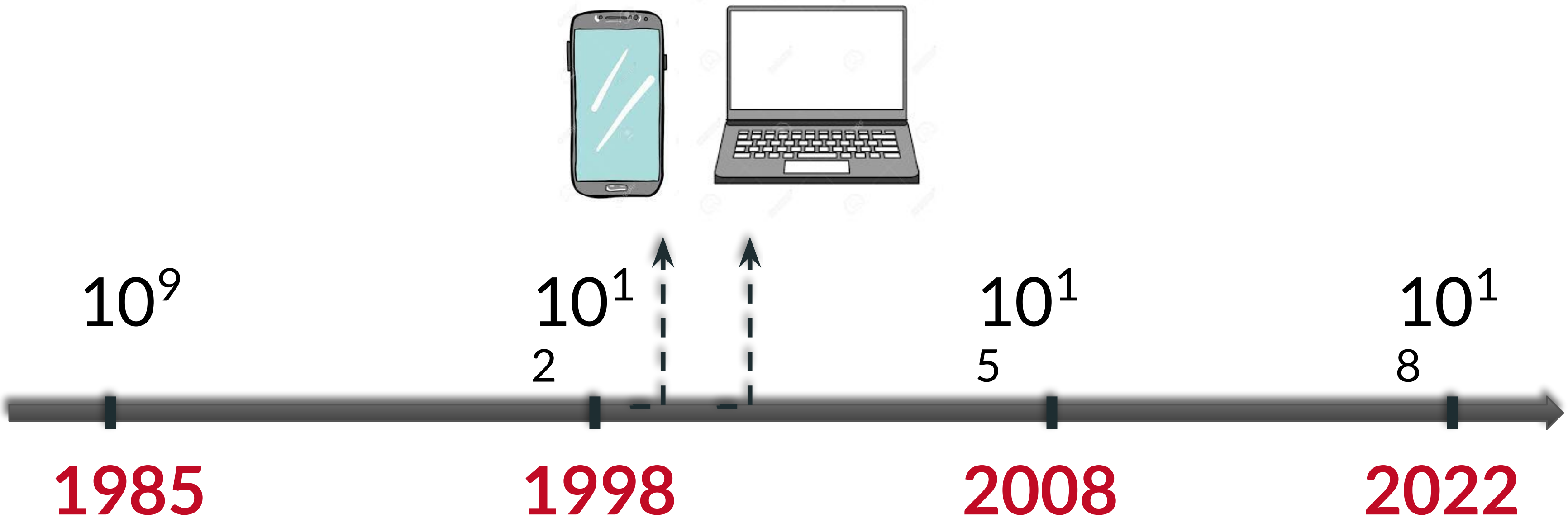
PAR CORE LAB

10^9	10^{12}	10^{15}	10^{18}
1985	**1998**	**2008**	**2022**

Thanks to Moore’s Law

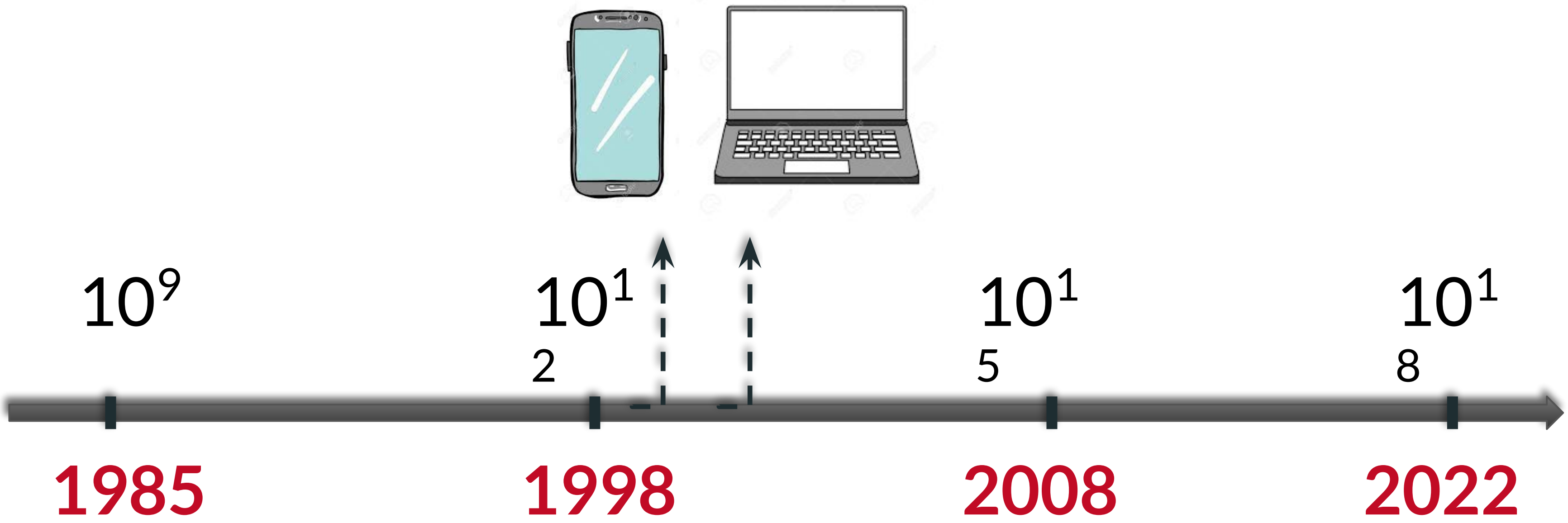


Thanks to Moore's Law



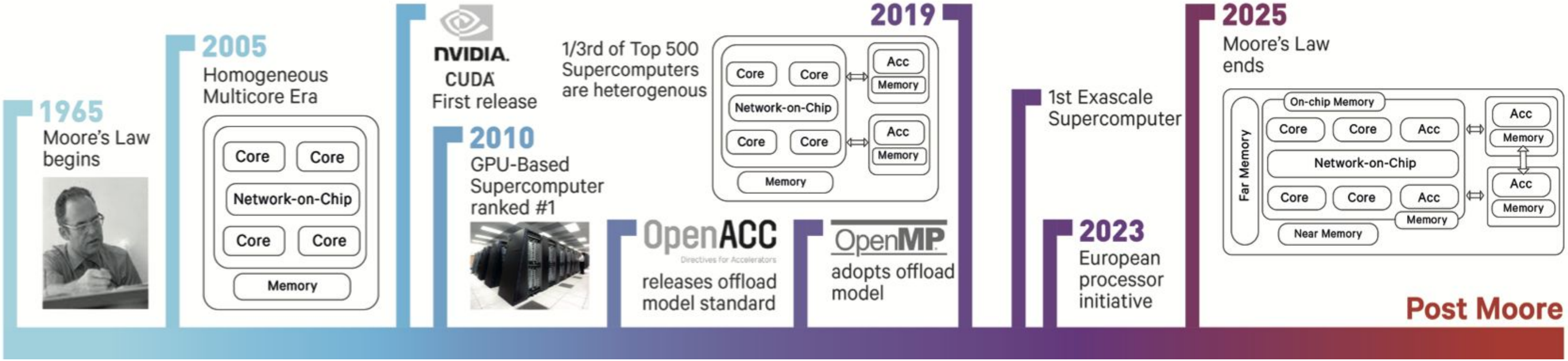
Observation 1: Supercomputing on a desktop

Thanks to Moore's Law

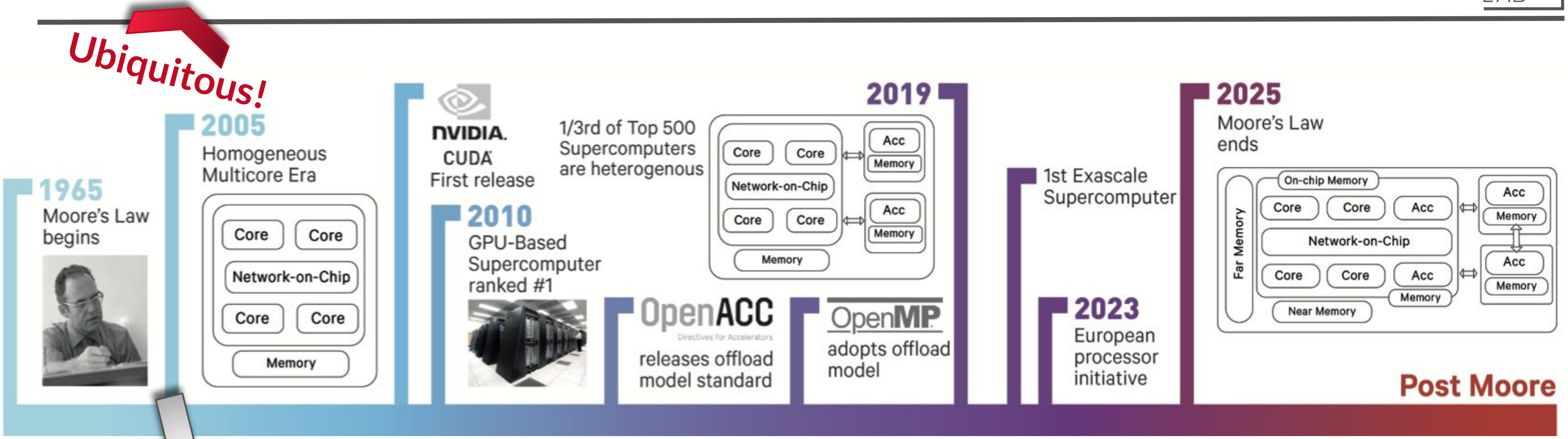


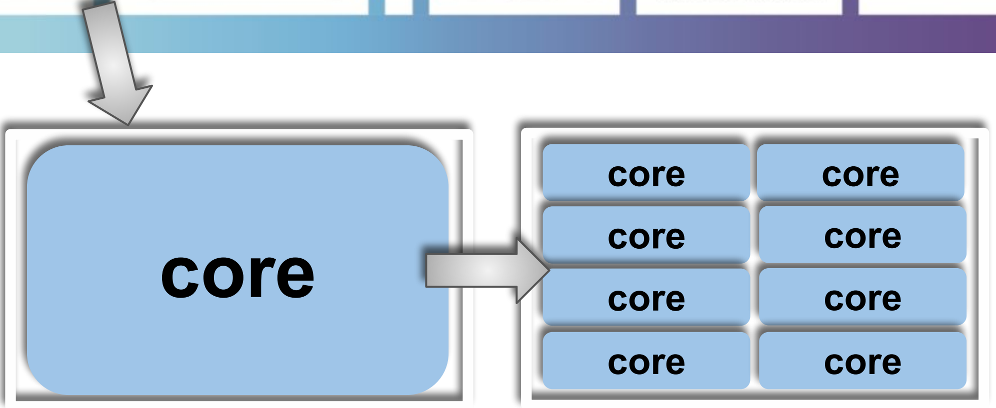
Observation 2: HPC software eventually will make it to the cell phones and laptops

A brief history of Moore’s Law



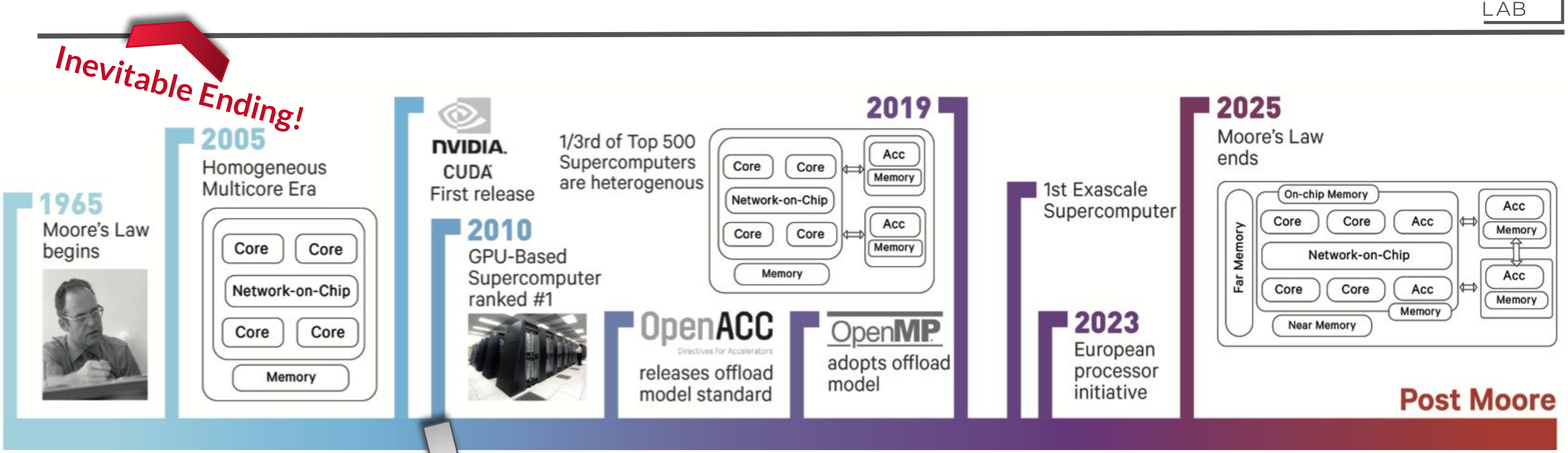
Trend #1 Multicore Architectures

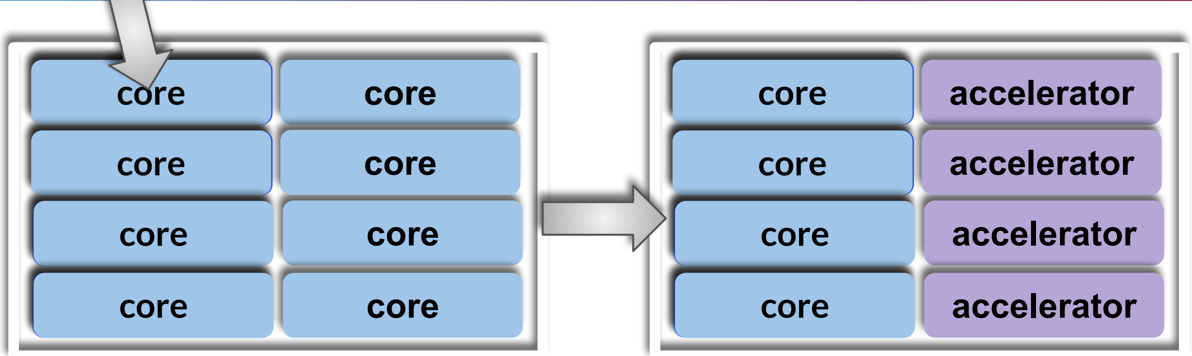




- Core count doesn't increase as much as we expected
- Teaching multicore programming is crucial

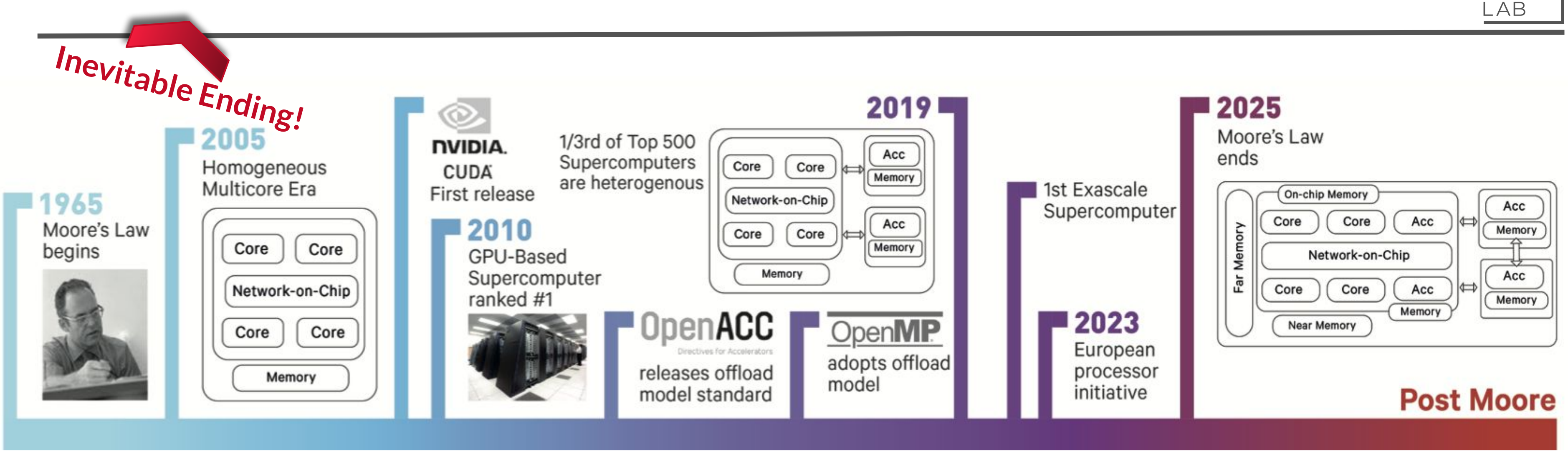
Trend #2 Heterogeneous Computing



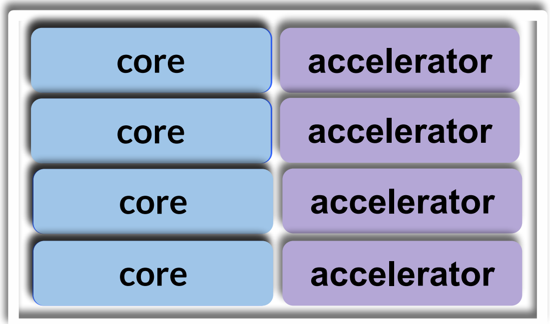


KOÇ ÜNİVERSİTESİ

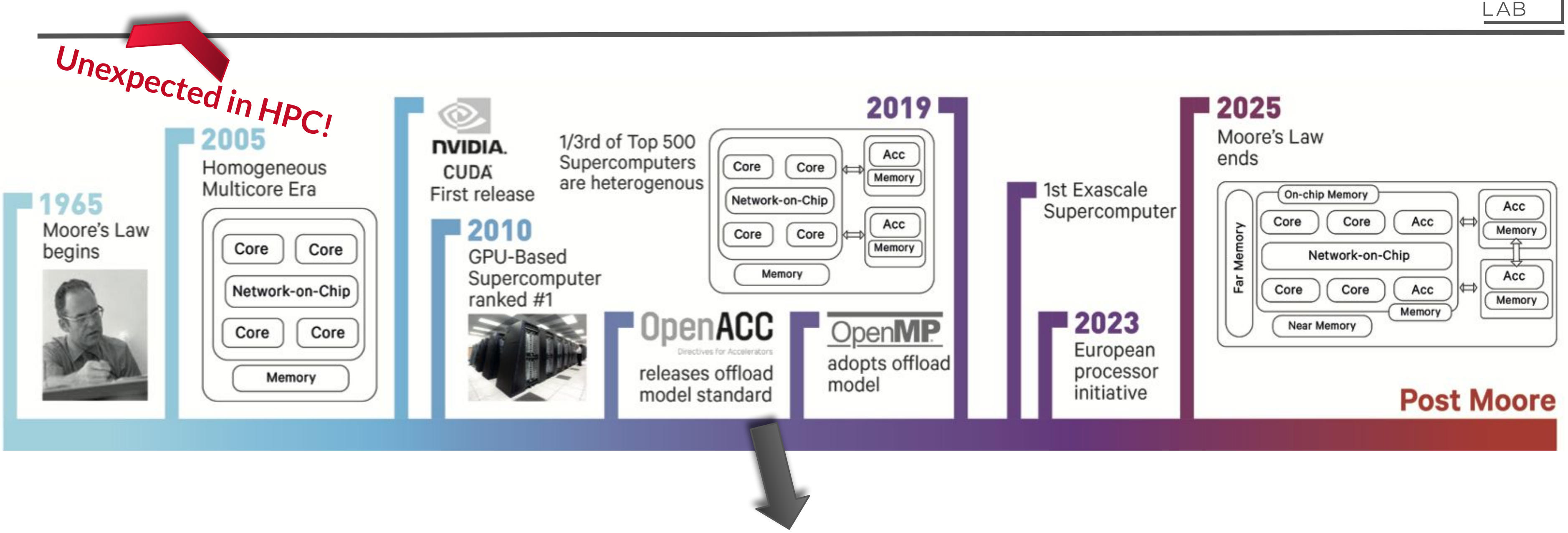
Trend #2 Heterogeneous Computing



- Several solutions how to program an accelerator
- Multi-GPU support is still not fully there
- Alternative but not the only solution



Trend #3 Emergence and Dominance of AI

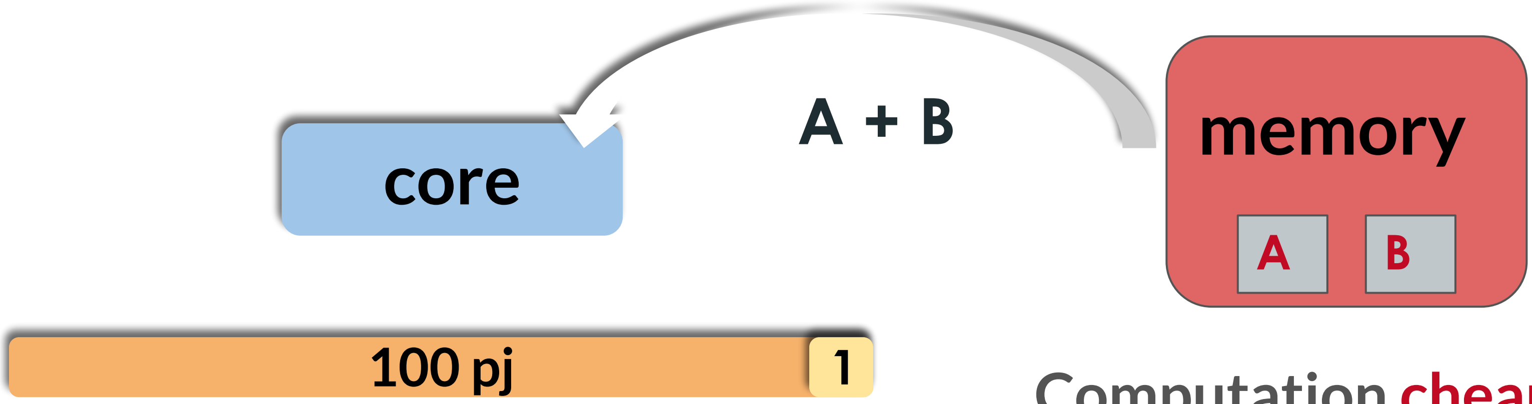


2016 onward, Parallel Training is The New Norm for DNN Training

Trend #4 Data movement cost

Unsolved!

Data Locality



Computation cheap!

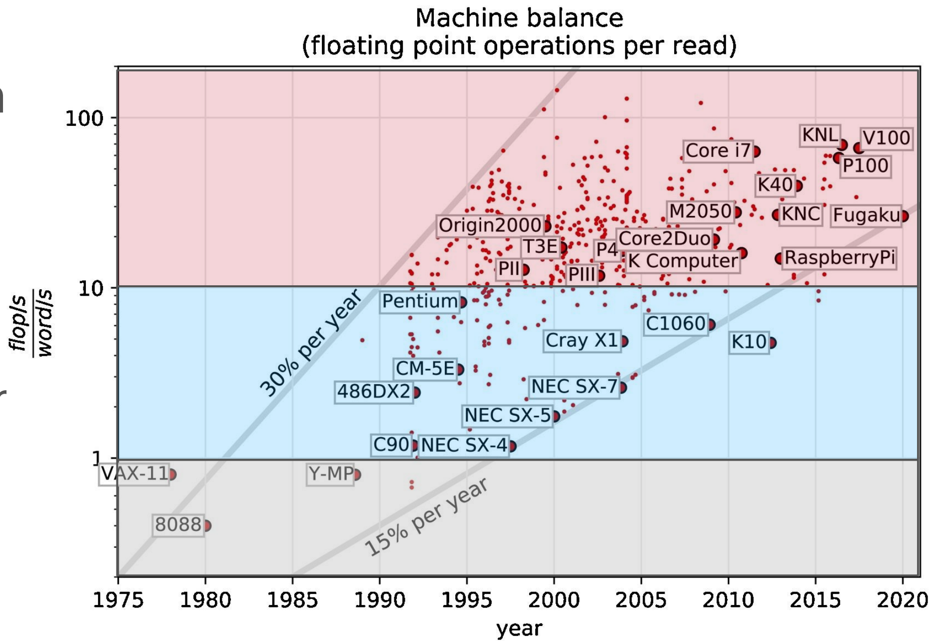
Data movement expensive!

[1]: Marc Horowitz, Computing's Energy Problem (and what we can do about it), ISSC 2014, plenary

Imbalance Machines

- Compute Speed/Memory Bandwidth Ratio: FLOPS/WORDS
- Today's systems are capable of doing 100 flops per word-transferred
- However applications do not perform that many operations

  => causing machine imbalance



Tomov, Translational
utational Science,

Leading to an increasing gap between data movement speeds and computation speeds.

	Top 500 System and Their HPCG Performance	Country	Cores	Rmax (PFlop/s)	TOP500 Rank	HPCG (PFlop/s)	% of Peak
1	Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu, RIKEN	Japan	7,630,848	442.01	2	16	3.00%
2	Frontier - HPE Cray EX235a, AMD EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE, DOE/SC/Oak Ridge	US	8,730,112	1,102.00	1	14.1	0.80%
3	LUMI - HPE Cray EX235a, AMD EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11 EuroHPC/CSC	Finland	2,220,288	309.1	3	3.41	0.80%
4	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM, DOE/SC/Oak Ridge	US	2,414,592	148.6	5	2.93	1.50%
5	Leonardo - Bull Sequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos, EuroHPC/CINECA	Italy	1,463,616	174.7	4	2.57	1.00%
6	Perlmutter - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, DOE/SC/LBNL/NERSC	US	761,856	70.87	8	1.91	2.00%
7	Sierra - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox, DOE/NNSA/LLNL	US	1,572,480	94.64	6	1.8	1.40%
8	Selene - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, NVIDIA Corporation	US	555,520	63.46	9	1.62	2.00%
9	JUWELS Booster Module - Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, Atos Forschungszentrum Juelich (FZJ)	Germany	449,280	44.12	12	1.28	1.80%
10	Dammam-7 - Cray CS-Storm, Xeon Gold 6248 20C 2.5GHz, NVIDIA Tesla V100 SXM2, InfiniBand HDR 100, HPE Saudi Aramco	Saudi Arabia	672,520	22.4	20	0.88	1.60%

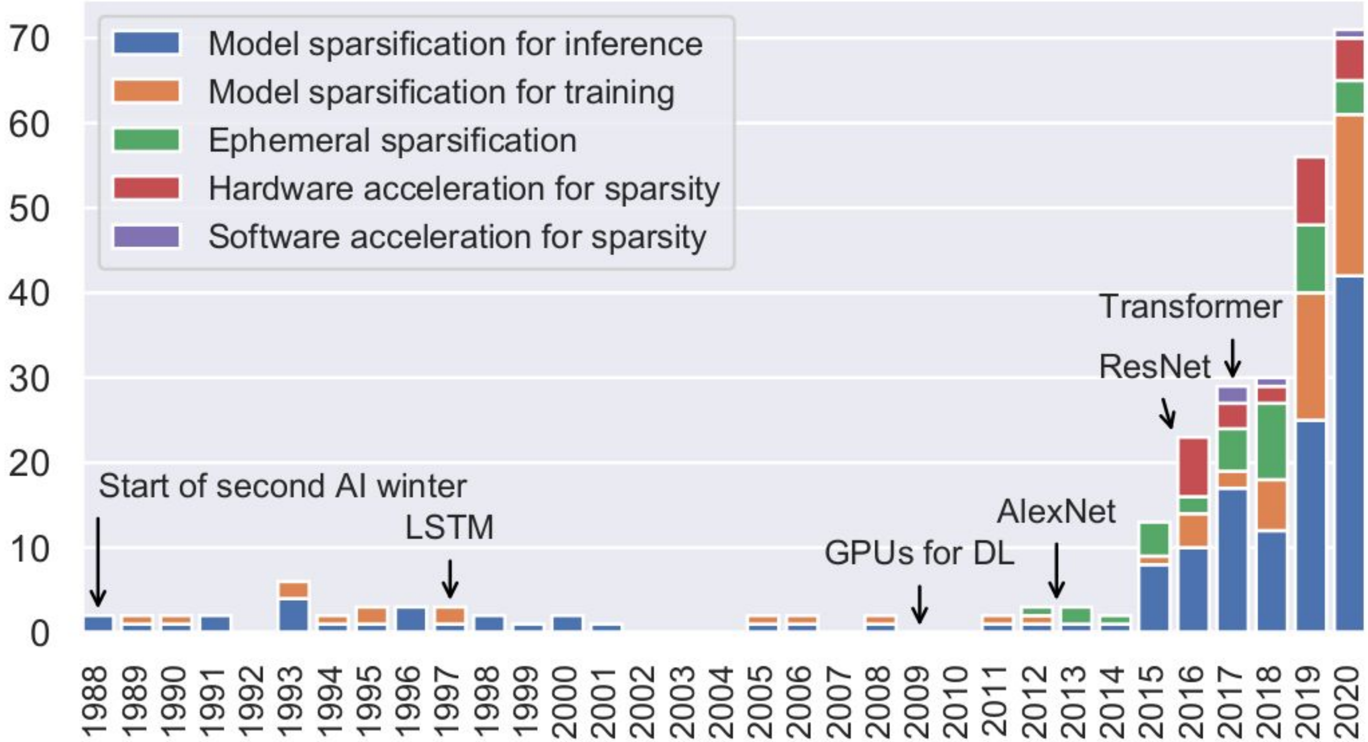
	Top 500 System and Their HPCG Performance	Country	Cores	Rmax (PFlop/s)	TOP500 Rank	HPCG (PFlop/s)	% of Peak
1	Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu, RIKEN	Japan	7,630,848	442.01	2	16	3.00%
2	Frontier - HPE Cray EX235a, AMD EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE, DOE/SC/Oak Ridge	US	8,730,112	1,102.00	1	14.1	0.80%
3	LUMI - HPE Cray EX235a, AMD EPYC 64C 2GHz, AMD Instinct MI250X, Sling…	Finland	2,220,288	309.1	3	3.41	0.80%
4	Sun… GV…						1.50%
5	Leo… A10… Eur…						1.00%
6	Perlmutter - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, DOE/SC/LBNL/NERSC	US	761,856	70.87	8	1.91	2.00%
7	Sierra - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mel… DOE/NNSA/LLNL	US	1,572,480	94.64	6	1.8	1.40%
8	Selene - NVIDIA DG… Mellanox HDR Infiniband, NVIDIA Corporation			…46	9	1.62	2.00%
9	JUWELS Booster Module - Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, Atos Forschungszentrum Juelich (FZJ)	Germany	449,280	44.12	12	1.28	1.80%
10	Dammam-7 - Cray CS-Storm, Xeon Gold 6248 20C 2.5GHz, NVIDIA Tesla V100 SXM2, InfiniBand HDR 100, HPE Saudi Aramco	Saudi Arabia	672,520	22.4	20	0.88	1.60%

- Clearly showing the machine imbalance
- Sparse computational kernels are harder to scale and optimize
- They are bounded by the memory/network bandwidth and latency

When will CG be able to sustain Exaflop performance?
Requires 100x performance improvement

KOÇ ÜNİVERSİTESİ

Dynamic Models: Sparsification



Sparsity in Deep Learning, T. Hoefler, https://arxiv.org/abs/2102.00554

Research Focus

PAR CORE LAB

- Data movement still dominates
- Hiding communication costs is becoming increasingly difficult but necessary
- At high levels of parallelism, synchronization becomes increasingly expensive.
- Stop counting FLOPS, count data movement

PADAL Workshop Series

PAR CORE LAB

- To discuss emerging approaches in data locality we organized a workshop at
  - Lugano, 2014
  - Berkeley, 2015
  - Kobe, 2016
  - Chicago, 2017
  - Bordeaux, 2019

- Major movement among code teams to implement data locality abstractions
  - because they are critically important for future codes.

IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS 1

Trends in Data Locality Abstractions for HPC Systems

Didem Unat, Anshu Dubey, Torsten Hoefler, John Shalf,
Mark Abraham, Mauro Bianco, Bradford L. Chamberlain, Romain Cledat, H. Carter Edwards, Hal Finkel,
Karl Fuerlinger, Frank Hannig, Emmanuel Jeannot, Amir Kamil, Jeff Keasler, Paul H J Kelly, Vitus Leung,
Hatem Ltaief, Naoya Maruyama, Chris J. Newburn, and Miquel Pericás

Abstract— The cost of data movement has always been an important concern in high performance computing (HPC) systems. It has now become the dominant factor in terms of both energy consumption and performance. Support for expression of data locality has been explored in the past, but those efforts have had only modest success in being adopted in HPC applications for various reasons. However, with the increasing complexity of the memory hierarchy and higher parallelism in emerging HPC systems, locality management has acquired a new urgency. Developers can no longer limit themselves to low-level solutions and ignore the potential for productivity and performance portability obtained by using locality abstractions. Fortunately, the trend emerging in recent literature on the topic alleviates many of the concerns that got in the way of their adoption by application developers. Data locality abstractions are available in the forms of libraries, data structures, languages and runtime systems; a common theme is increasing productivity without sacrificing performance. This paper examines these trends and identifies commonalities that can combine various locality concepts to develop a comprehensive approach to expressing and managing data locality on future large-scale high-performance computing systems.

Index Terms—Data locality, programming abstractions, high-performance computing, data layout, locality-aware runtimes

KOÇ ÜNİVERSİTESİ

Sparse Computation

Alternative Model of Execution

Data Locality Tools

An Optimization and Co-design Framework for Sparse Computation

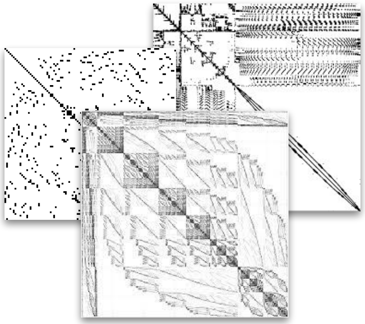
Didem Unat

Project Coordinator, Koç University

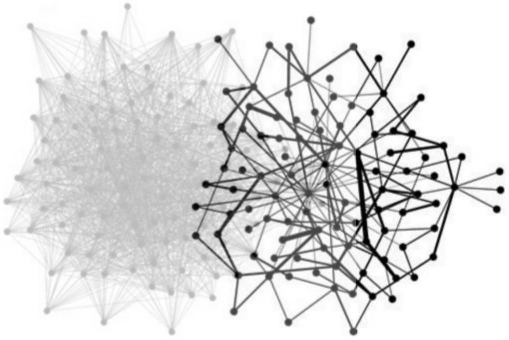
This project has received funding from the European High-Performance Computing Joint Undertaking under grant agreement No. 956213.

Feature extraction of sparse computation

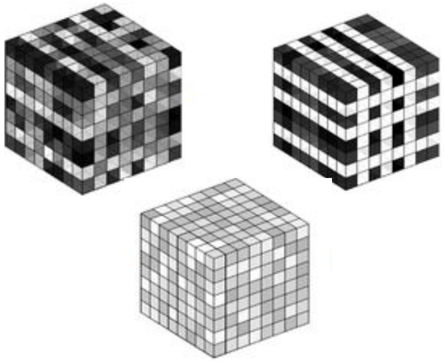
SPARSE MATRICES



GRAPHS

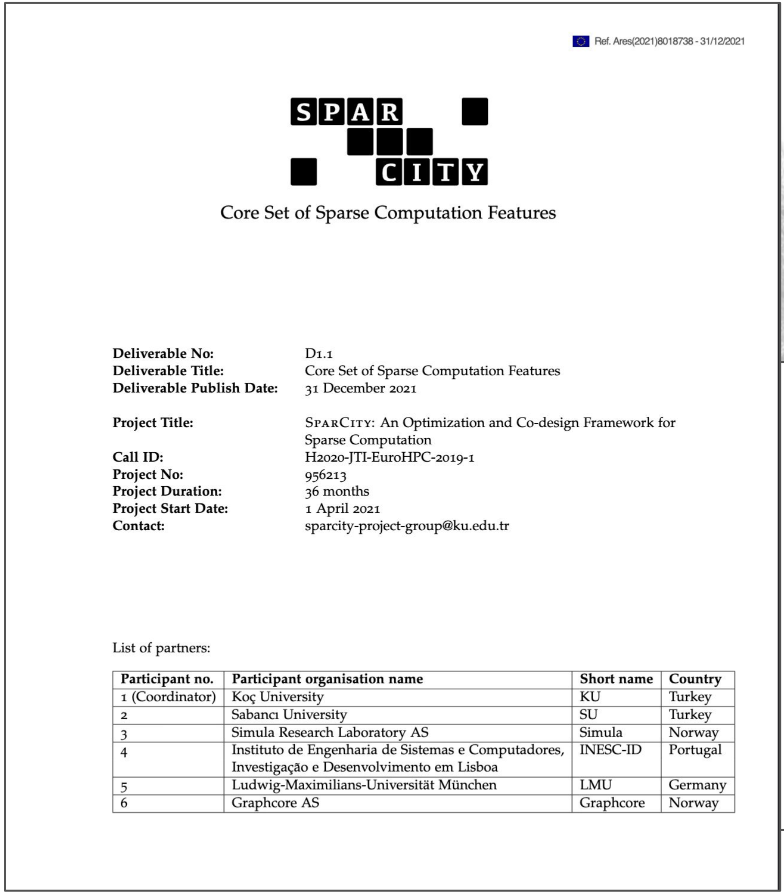


SPARSE TENSORS



- **Efficient sparse computing depends on statistical features that describe the sparsity pattern**
- **More than 100 extracted features** integrated in one large and practical set of sparse features
- Used as inputs for automated sparse format and kernel selection ML-based approaches
- Low-overhead extraction methods for **integration in the SparseBase framework**

Feature extraction of sparse computation


Ref. Ares(2021)8018738 - 31/12/2021

SPAR CITY

Core Set of Sparse Computation Features

Deliverable No: D1.1
Deliverable Title: Core Set of Sparse Computation Features
Deliverable Publish Date: 31 December 2021

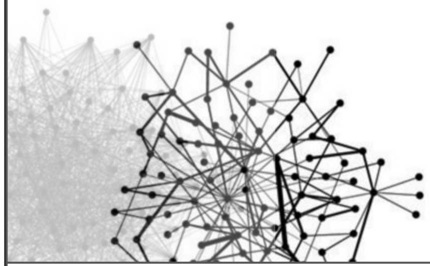
Project Title: SparCity: An Optimization and Co-design Framework for Sparse Computation

Call ID: H2020-JTI-EuroHPC-2019-1
Project No: 956213
Project Duration: 36 months
Project Start Date: 1 April 2021
Contact: sparcity-project-group@ku.edu.tr

List of partners:

Participant no.	Participant organisation name	Short name	Country
1 (Coordinator)	Koç University	KU	Turkey
2	Sabancı University	SU	Turkey
3	Simula Research Laboratory AS	Simula	Norway
4	Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa	INESC-ID	Portugal
5	Ludwig-Maximilians-Universität München	LMU	Germany
6	Graphcore AS	Graphcore	Norway

GRAPHS



SPARSE TENSORS

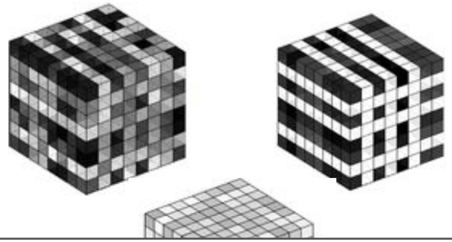
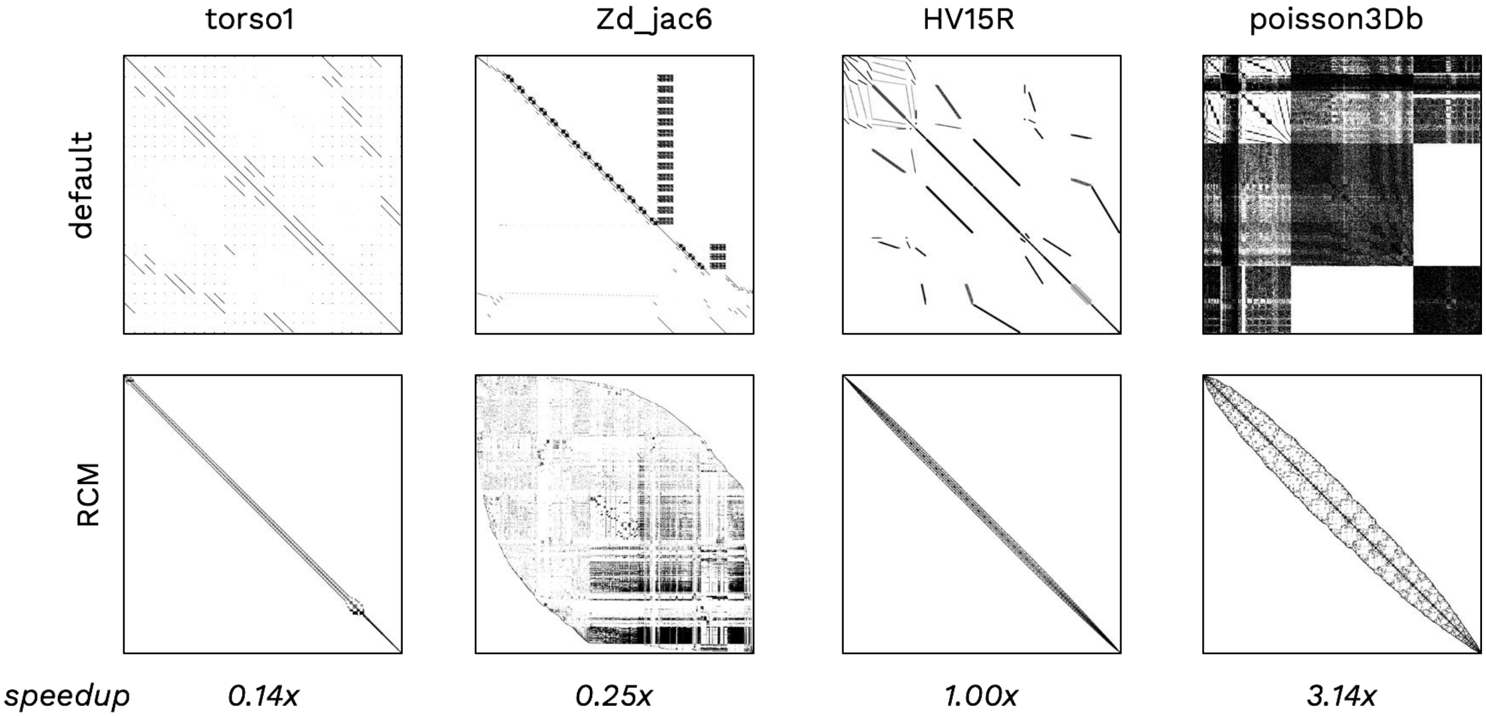


Table 6 *Sparse tensor features. *Due to the complex formula, the formula is omitted but described in the text.*

Feature	Description	Formula
size_m	Tensor mode size (in mode m)	I_m
fiberCnt	Number of fibers	$\lvert\mathcal{F}\rvert = \sum \lvert\mathcal{F}_m\rvert$
sliceCnt	Number of slices	$\lvert\mathcal{S}\rvert = \sum \lvert\mathcal{S}_{k,\ell}\rvert$
fiberRatio	The ratio of fibers	$\lvert\mathcal{F}\rvert/\mathtt{min}(\lvert\mathcal{F}_m\rvert)$
sliceRatio	The ratio of slices	$\lvert S\rvert/\mathtt{min}(\lvert\mathcal{S}_{k,\ell}\rvert)$
nnz	Number of nonzeros	$\mathtt{nzz}(\mathcal{X})$
density	Density of nnz in the tensor	$\mathtt{nzz}(\mathcal{X})/\prod I_m$
avgNnzPerSlice	Average nnz per slice	$\sum \mathtt{nnz}(\mathcal{S}_{k,\ell})/\lvert\mathcal{S}\rvert$
maxNnzPerSlice	Maximum nnz per slice	$\mathtt{max}(\mathtt{nnz}(\mathcal{S}_{k,\ell}))$
minNnzPerSlice	Minimum nnz per slice	$\mathtt{min}(\mathtt{nnz}(\mathcal{S}_{k,\ell}))$
adjNnzPerSlice	Average nnz difference of adjacent slices	*
devNnzPerSlice	The deviation of nnz per slice	*

Reordering for better caching

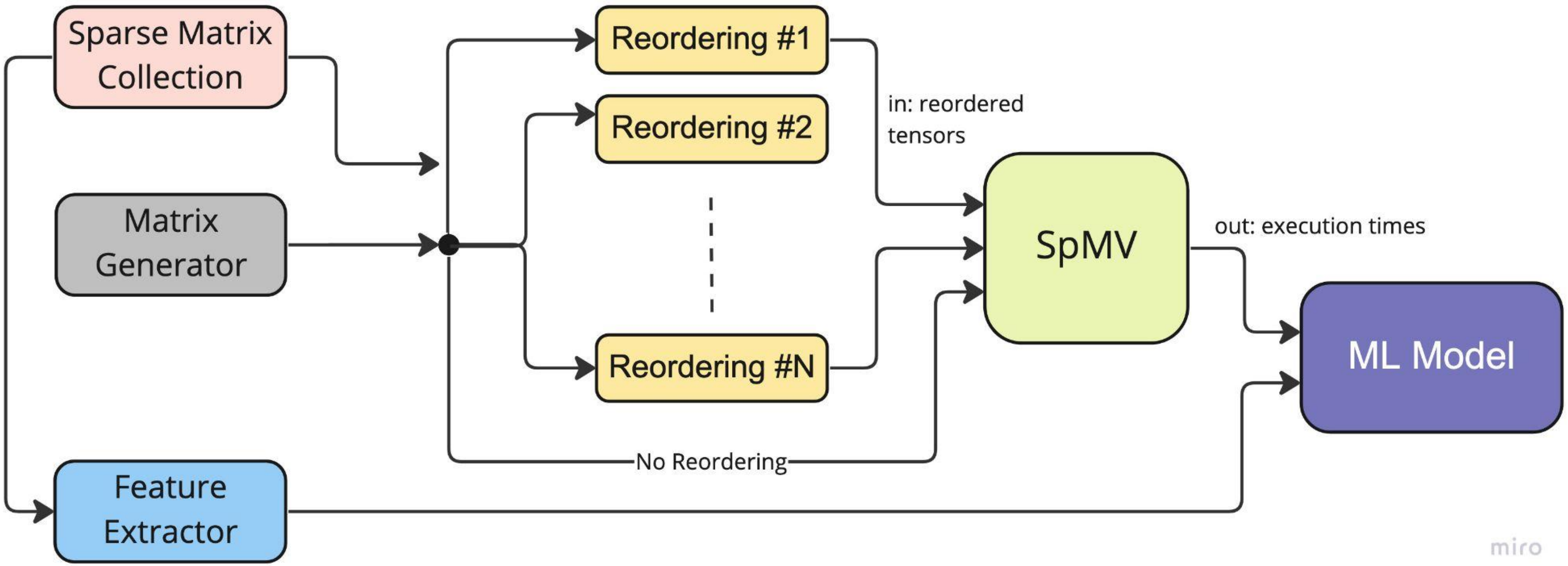
Reverse Cuthill-McKee (RCM): heuristic bandwidth minimization for SpMV



Slide credit: Johannes Langguth (Simula)

Reordering Algorithm Selection

PAR CORE LAB

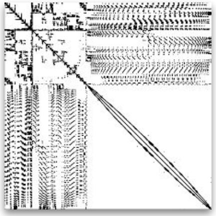


In collaboration with SparCity partners

SparseBase: Pre-processing Base for Sparse Computation

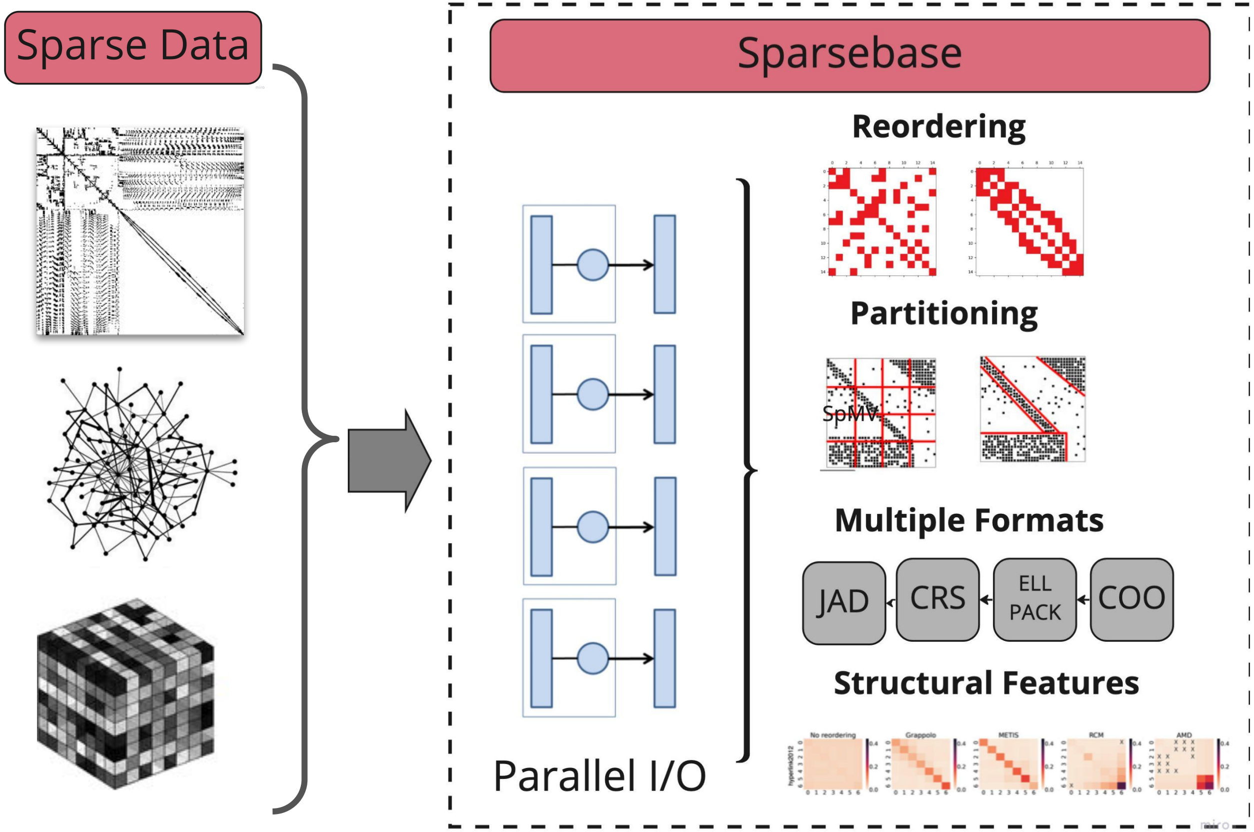
https://github.com/sparcityeu/sparsebase

Sparse Data



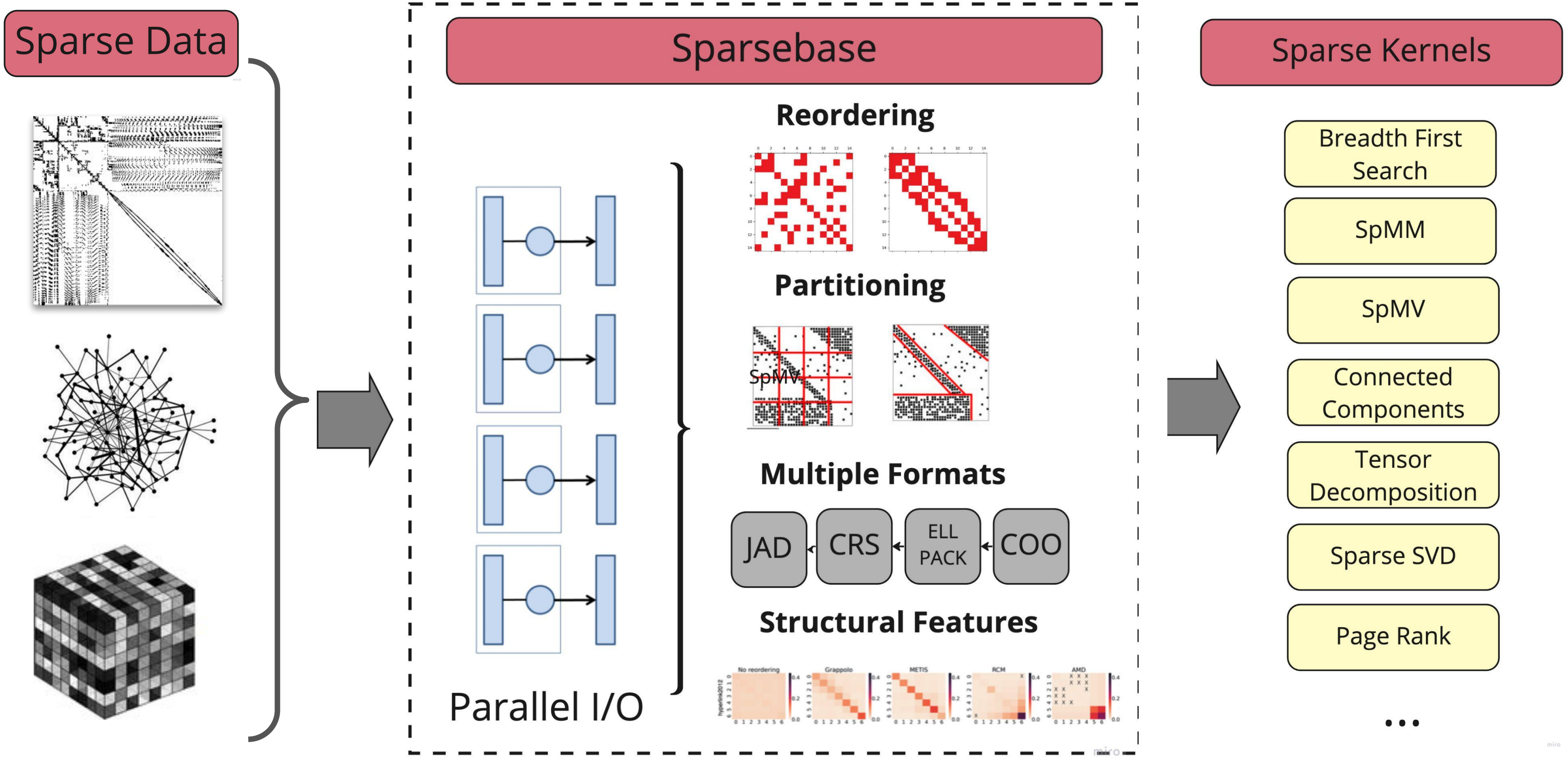
SparseBase: Pre-processing Base for Sparse Computation

https://github.com/sparcityeu/sparsebase



SparseBase: Pre-processing Base for Sparse Computation

https://github.com/sparcityeu/sparsebase



KOÇ ÜNİVERSİTESİ

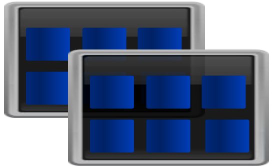
CPU-free Execution

European Research Council
Established by the European Commission

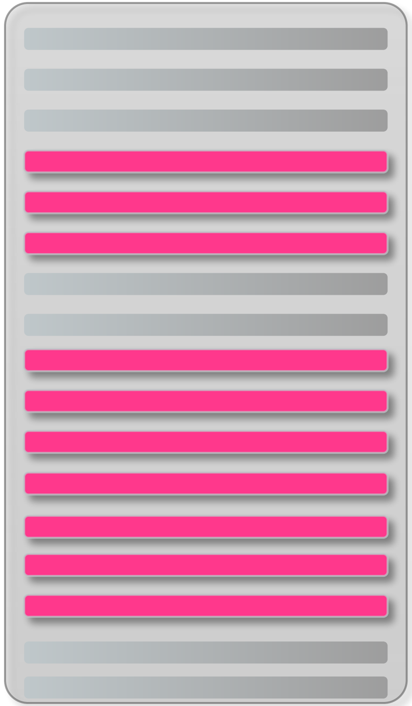
This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 949587).

Traditional GPU Execution Model

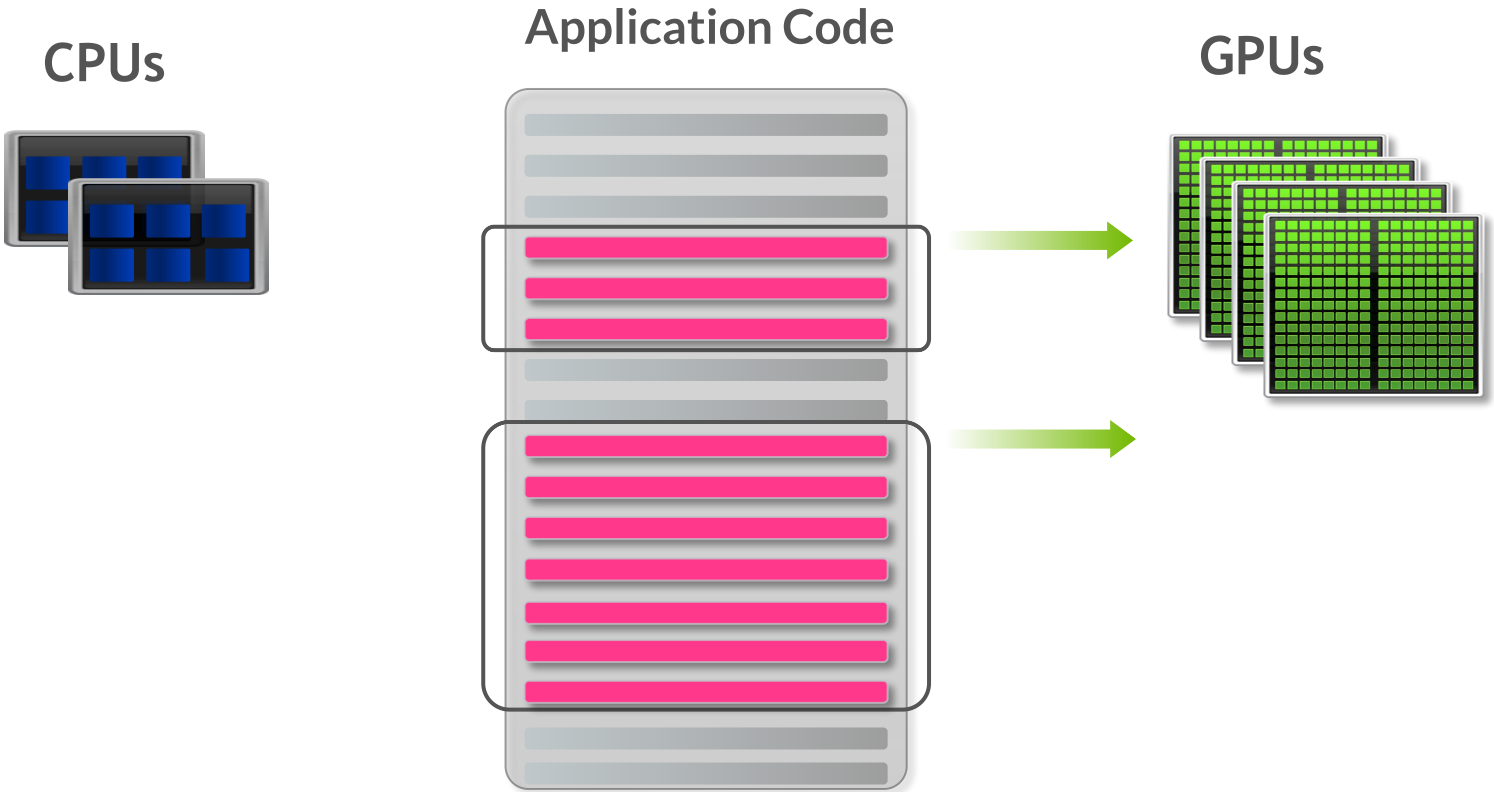
CPUs



Application Code



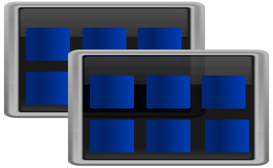
Traditional GPU Execution Model



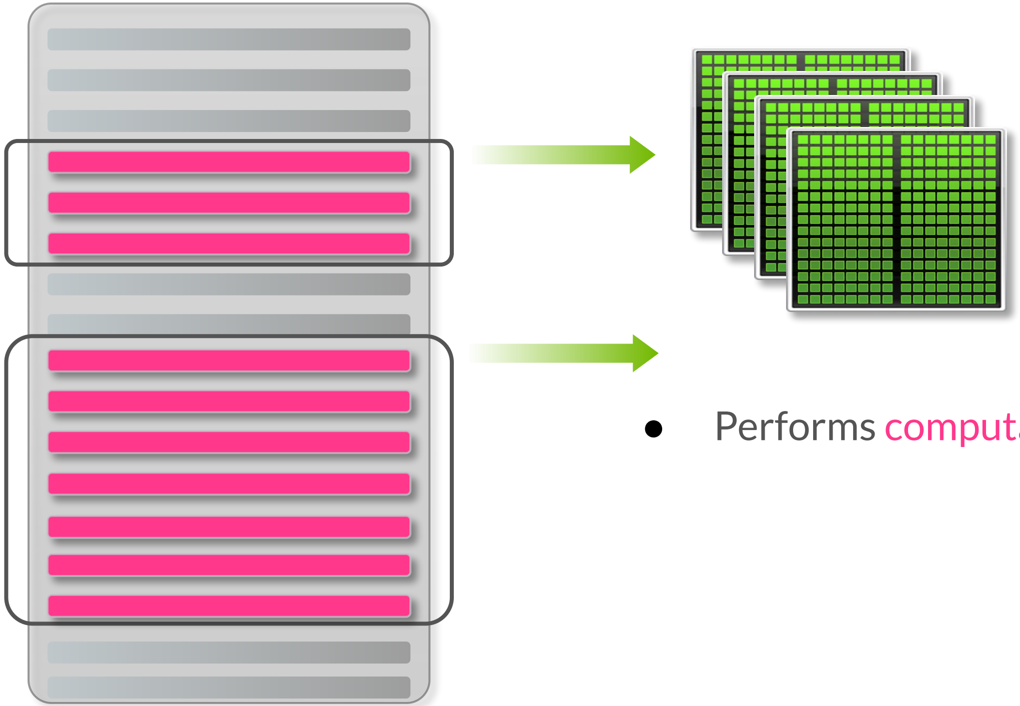
Traditional GPU Execution Model

Application Code

CPUs



- Launches compute kernels
- Issues communication calls
- Overlaps communication with compute
- Acts as synchronizer both within and across devices

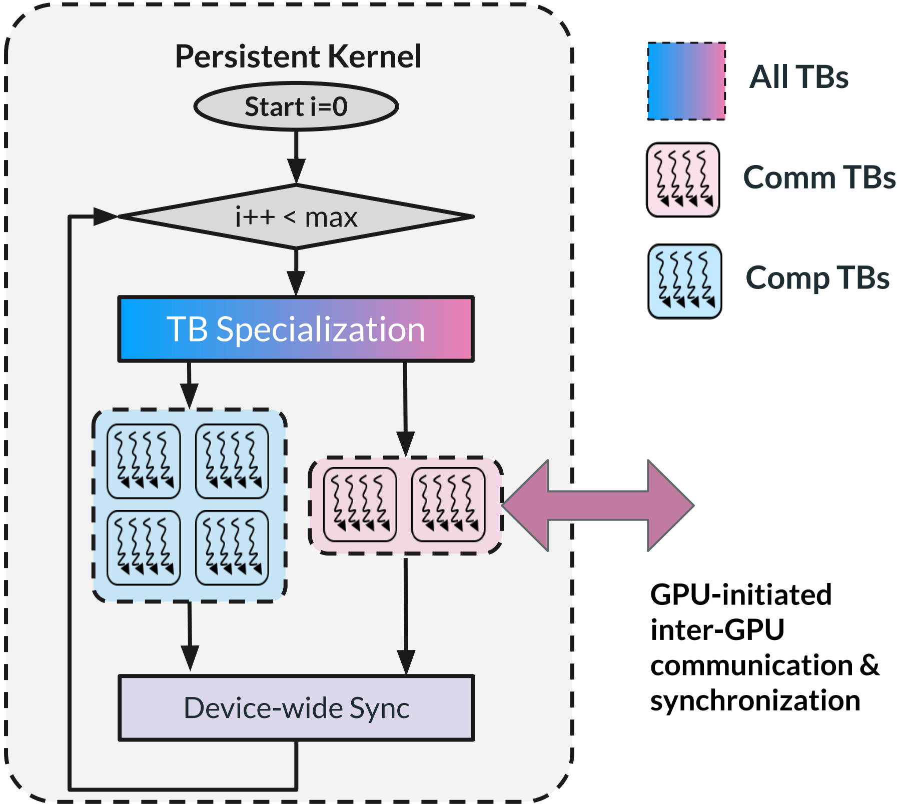


GPUs

- Performs computation

CPU-free execution model

- **Persistent kernels**
  - Long running kernels
  - Time loop on the device
- **TB specialization**
  - Spare some TBs to comm
  - Rest for computation
- **GPU-initiated data movement**
  - Peer to peer communication
  - Issue calls within the device
- **Device-side synchronization**
  - Device -wide barriers
  - Across device sync



Data Locality Tools

Data-Centric Profiling Tools

ComDetective

- **"ComDetective: A Lightweight Communication Detection Tool for Threads"**, IEEE/ACM Supercomputing Conference, **Best Student Paper and Best Paper finalist for SC19**.

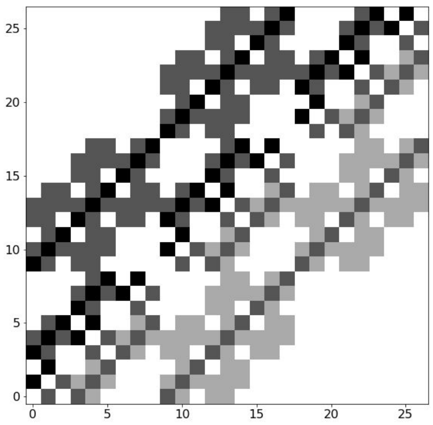
ReuseTracker

- "**ReuseTracker: Fast Yet Accurate Multicore Reuse Distance Analyzer**", ACM TACO and HiPEAC Conference, June 2022

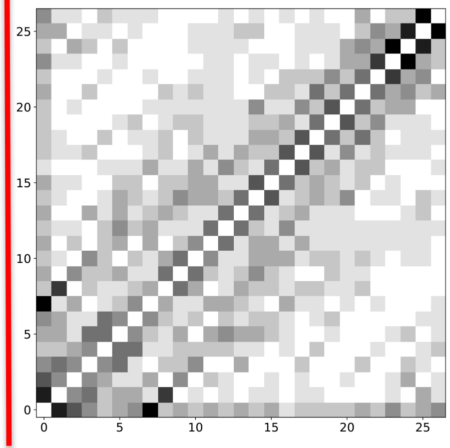
AMD vs Intel

- ***"Precise Event Sampling on AMD vs Intel: Quantitative and Qualitative Comparison"****, IEEE TPDS, 2023. Under minor revision.*

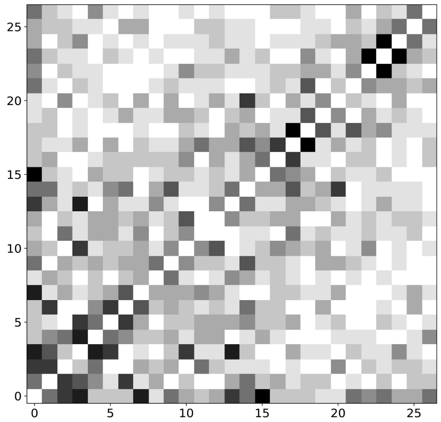
ComDetective



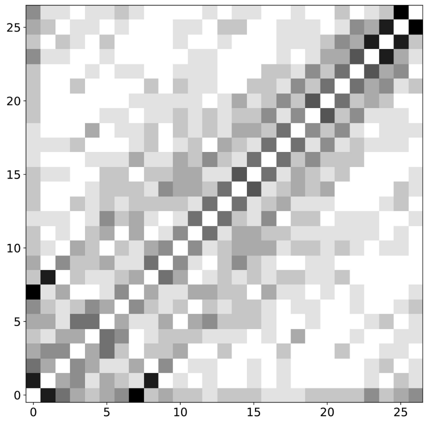
MPI communication matrix



communication matrix



true sharing matrix



false sharing matrix

- In addition to communication matrix, ComDetective also produces true sharing and false sharing matrices
- It took only **1.28x** performance and **1.11x** memory footprint overhead to generate these matrices with ComDetective

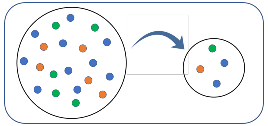
ReuseTracker

Fast

introducing low time overhead

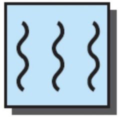
Accurate

Verified against a set of microbenchmarks with known ground truths



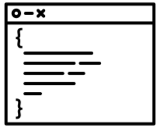
Sampling based

Uses ready-available hardware events in commodity CPUs



Multithreads

Profiles reuse distance in **private** and **shared** caches by also detecting cache line invalidations



Code line attribution

Attributes uses and reuses to source code lines

Open source

Usage

- ComDetective
  - Guiding thread-to-core binding
  - Guiding code refactoring to remove false sharing

- ReuseTracker
  - Loop transformation that should be implemented to improve locality in local data caches
  - Decide whether to activate hardware-level optimizations such as adjacent cache line prefetch to reduce high latency memory accesses

https://github.com/ParCoreLab/ParCoreTools

Outlook

Future of computing is exciting

PAR
CORE
LAB

- Parallel systems are here to stay
  - Homogeneous and heterogeneous systems

- Reducing data movement and synchronization overheads are crucial

- Post-Moore's Era
  - Requires more interaction with vendors and AI community to solve the data movement problems
  - HPC can help AI, AI can help HPC

- Plenty of research opportunities
  - We are hiring !!!

Supercomputers mean super power.

By developing software for supercomputers, one can control that super power.

Call for Collaboration

PAR CORE LAB

For disaster management and large-scale earthquake simulation

With the help of HPC

- Ground motion prediction
- Spectral acceleration measurements with high resolution
- Detailed seismic hazard maps
- AI guided building design

Infrastructure support, simulation/modeling, parallelization, domain expertise

Earth’s crust has gone a deformation of 3-4 meters along a line of approximately 400 kms (250 miles)

https://www.nytimes.com/interactive/2023/02/10/world/middleeast/kahramanmaras-turkey-earthquake-damage.html?smid=tw-nytimesworld&smtyp=cur

Thank you!

https://parcorelab.ku.edu.tr/

dunat@ku.edu.tr

KOÇ ÜNİVERSİTESİ

PAR CORE LAB