Towards sustainable AI: Performance optimization and workload consolidation strategies

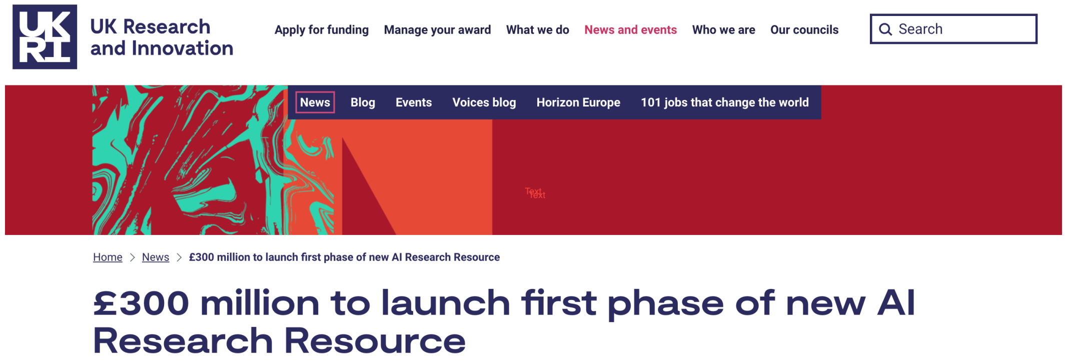
Nikela Papadopoulou

University of Glasgow

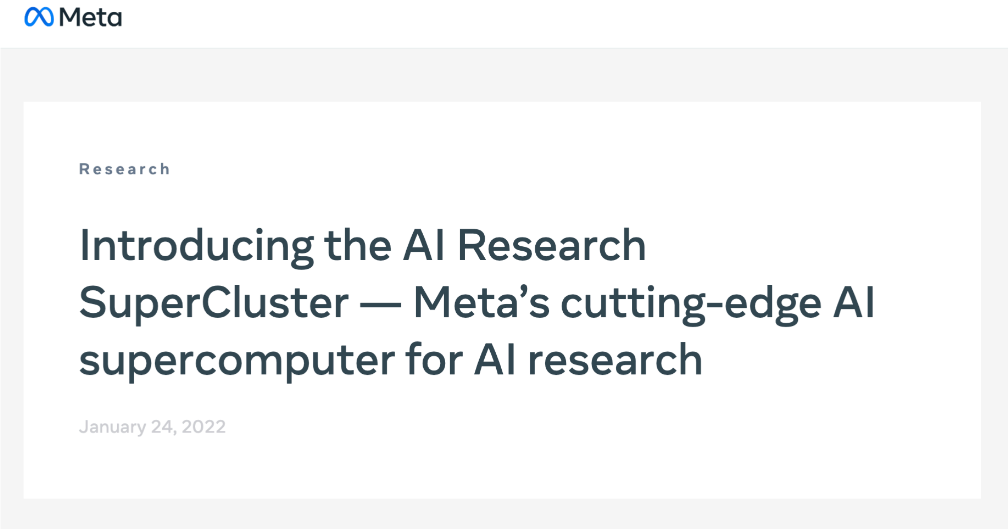
Low Carbon and Sustainable Computing Seminar

February 8, 2024

AI and HPC are converging fast


UK Research and Innovation
Apply for funding
Manage your award
What we do
News and events
Who we are
Our councils
Search
News
Blog
Events
Voices blog
Horizon Europe
101 jobs that change the world
Home > News > £300 million to launch first phase of new AI Research Resource

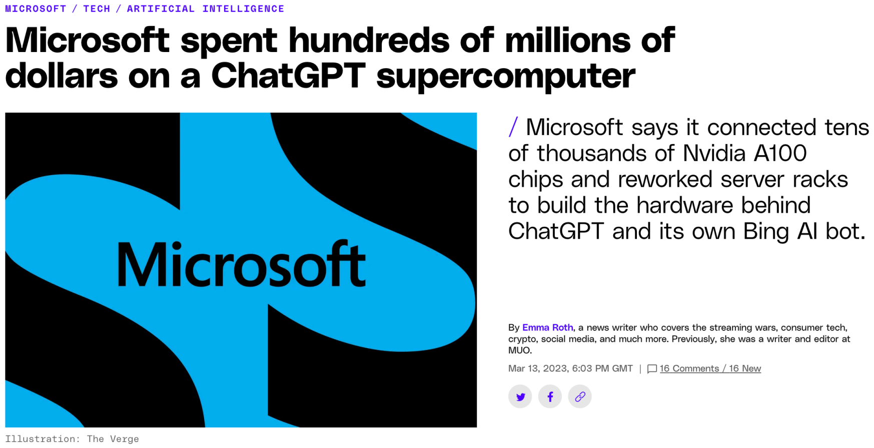
£300 million to launch first phase of new AI Research Resource



Research

Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research

January 24, 2022



Microsoft spent hundreds of millions of dollars on a ChatGPT supercomputer

/ Microsoft says it connected tens of thousands of Nvidia A100 chips and reworked server racks to build the hardware behind ChatGPT and its own Bing AI bot.

By Emma Roth, a news writer who covers the streaming wars, consumer tech, crypto, social media, and much more. Previously, she was a writer and editor at MUO.

Mar 13, 2023, 6:03 PM GMT | 16 Comments / 16 New

Illustration: The Verge

AI and HPC are converging fast

AI requirements are fulfilled by *HPC offerings*

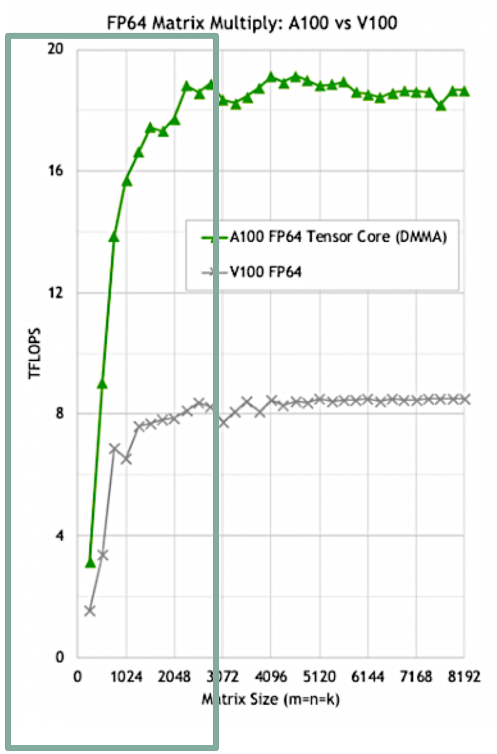
1. Intensive tensor computations → *Optimized dense linear algebra*
2. Extreme numbers of parameters → *Distribution, Parallelization, Communication*
3. Multiple end-users → *Multi-tenancy (in the cloud)*
4. Extreme data needs → *Fast storage*

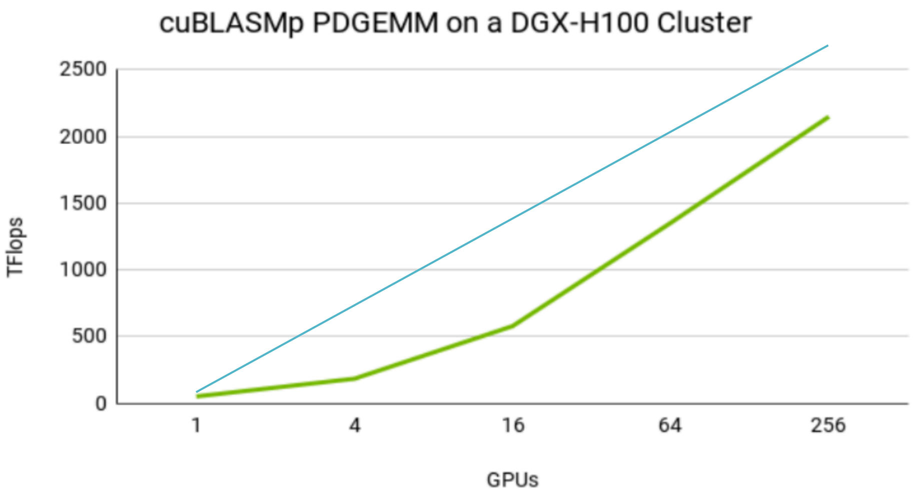
Case #1

Matrix Multiplication on GPUs

AI is dense matrix-matrix multiplication

A trend driving/driven by GPUs



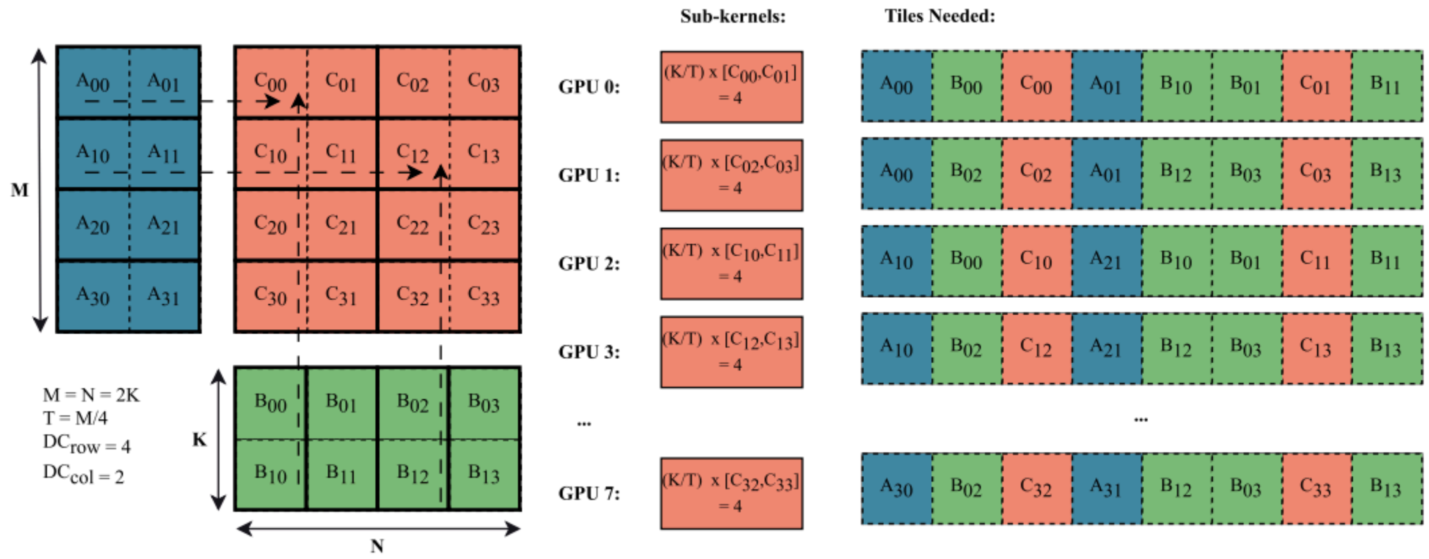


Weak scaling of cuBLASMp distributed double precision GEMM. M,N,K = 55k per GPU

Source: https://developer.nvidia.com/cublas

Matrix Multiplication on GPUs

State of Practice: 2D decomposition for matrix multiplication


A_{00} A_{01}
A_{10} A_{11}
A_{20} A_{21}
A_{30} A_{31}
M
C_{00} C_{01} C_{02} C_{03}
C_{10} C_{11} C_{12} C_{13}
C_{20} C_{21} C_{22} C_{23}
C_{30} C_{31} C_{32} C_{33}
B_{00} B_{01} B_{02} B_{03}
B_{10} B_{11} B_{12} B_{13}
K
N
$M = N = 2K$
$T = M/4$
$DC_{row} = 4$
$DC_{col} = 2$

	Sub-kernels:	Tiles Needed:							
GPU 0:	(K/T) x $[C_{00}, C_{01}]$ = 4	A_{00}	B_{00}	C_{00}	A_{01}	B_{10}	B_{01}	C_{01}	B_{11}
GPU 1:	(K/T) x $[C_{02}, C_{03}]$ = 4	A_{00}	B_{02}	C_{02}	A_{01}	B_{12}	B_{03}	C_{03}	B_{13}
GPU 2:	(K/T) x $[C_{10}, C_{11}]$ = 4	A_{10}	B_{00}	C_{10}	A_{21}	B_{10}	B_{01}	C_{11}	B_{11}
GPU 3:	(K/T) x $[C_{12}, C_{13}]$ = 4	A_{10}	B_{02}	C_{12}	A_{21}	B_{12}	B_{03}	C_{13}	B_{13}
...		...							
GPU 7:	(K/T) x $[C_{32}, C_{33}]$ = 4	A_{30}	B_{02}	C_{32}	A_{31}	B_{12}	B_{03}	C_{33}	B_{13}

Challenges:

- Decomposition – tile size selection
- Communication overlap, Communication avoidance, Communication routing
- Load balancing

[P. Anastasiadis, et al., *CoCoPeLia: Communication-Computation Overlap Prediction for Efficient Linear Algebra on GPUs,* ISPASS 2021]

[P. Anastasiadis, et al., *PARALiA : A Performance Aware Runtime for Auto-tuning Linear Algebra on heterogeneous systems,* ACM TACO 2023]

Matrix Multiplication on GPUs

Can we autotune dense linear algebra libraries on multiple GPUs?

- Yes – BLASX, XKBLAS, cuBLASXt, cuBLASLt

What is missing?

- Data placement
- Effective computation-communication overlap
- Topology awareness
- Device selection for energy efficiency

[P. Anastasiadis, et al., *CoCoPeLia: Communication-Computation Overlap Prediction for Efficient Linear Algebra on GPUs,* ISPASS 2021]

[P. Anastasiadis, et al., *PARALiA : A Performance Aware Runtime for Auto-tuning Linear Algebra on heterogeneous systems,* ACM TACO 2023]

Matrix Multiplication on GPUs

CoCoPeLIA: BLAS autotuner on a single GPU

for effective Computation-Communication overlap on the CPU-GPU link

1. Design performance models
   - Data location awareness
   - Bidirectional overlap
   - Data reuse
2. Autotune the tile size

Performance improvements (compared to SoTA):

- 5-15% on average
- 16-33% when the problem initially is on the CPU

[P. Anastasiadis, et al., *CoCoPeLia: Communication-Computation Overlap Prediction for Efficient Linear Algebra on GPUs,* ISPASS 2021]

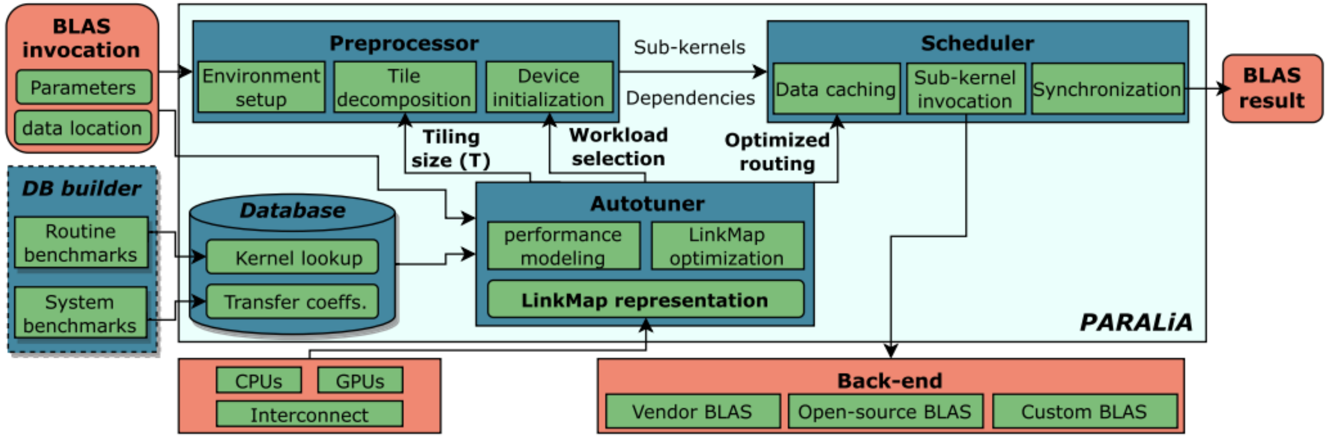
Matrix Multiplication on GPUs

PARALiA: a runtime for autotuning BLAS on multi-GPU systems

for Topology-aware Communication and Computation-Communication overlap

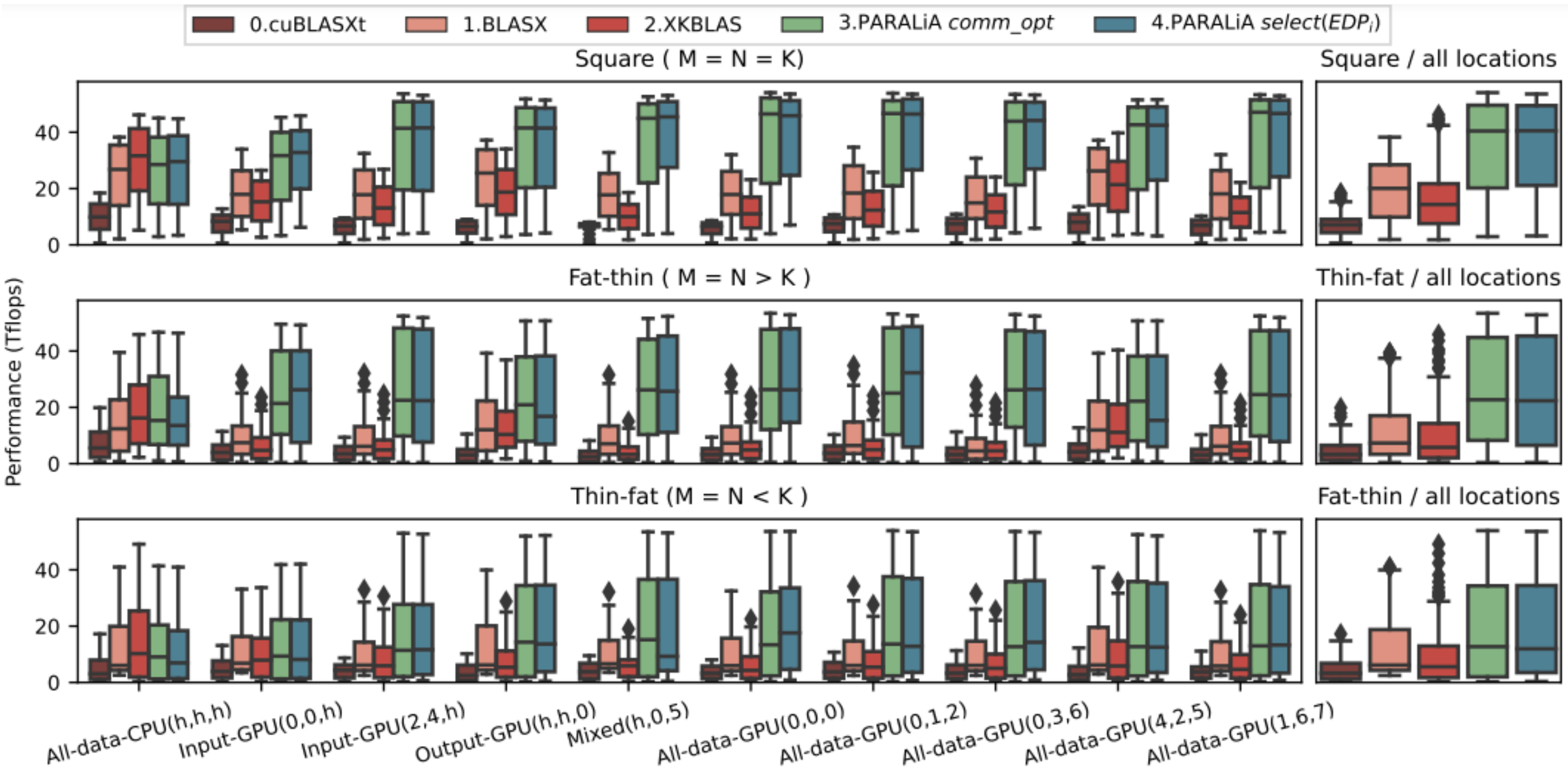
1. Design performance models
2. Design an abstract topology representation
3. Optimize communication paths
4. Autotune with optimal tile size
5. Schedule tiles

   according to optimization target



[P. Anastasiadis, et al., *PARALiA : A Performance Aware Runtime for Auto-tuning Linear Algebra on heterogeneous systems,* ACM TACO 2023]

Matrix Multiplication on GPUs



Performance improvements (average-compared to SoTA):

- 1.7x performance
- 2.5x energy efficiency

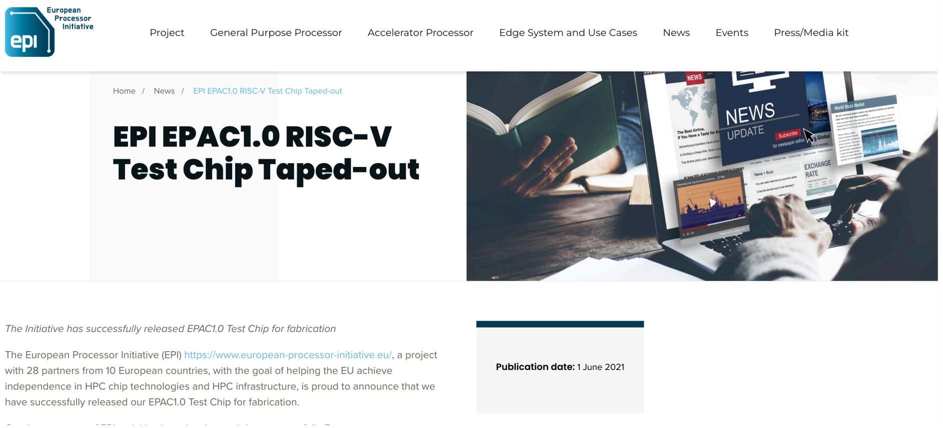
[P. Anastasiadis, et al., *PARALiA : A Performance Aware Runtime for Auto-tuning Linear Algebra on heterogeneous systems,* ACM TACO 2023]

Case #2

Matrix Multiplication on emerging architectures

AI is dense matrix-matrix multiplication

Are GPUs all we have?



European Processor Initiative
epi
Project
General Purpose Processor
Accelerator Processor
Edge System and Use Cases
News
Events
Press/Media kit
Home / News / EPI EPAC1.0 RISC-V Test Chip Taped-out

EPI EPAC1.0 RISC-V Test Chip Taped-out

The Initiative has successfully released EPAC1.0 Test Chip for fabrication

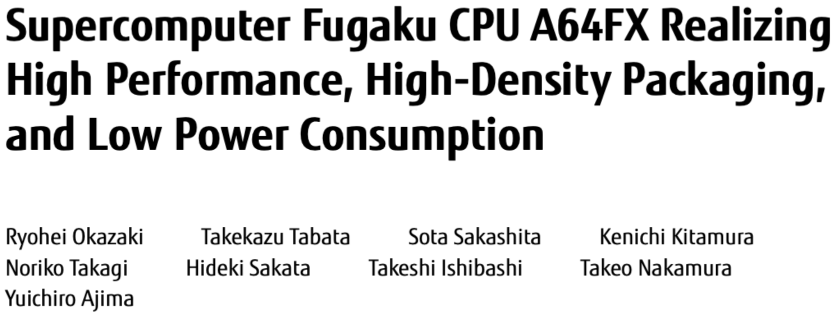
The European Processor Initiative (EPI) https://www.european-processor-initiative.eu/, a project with 28 partners from 10 European countries, with the goal of helping the EU achieve independence in HPC chip technologies and HPC infrastructure, is proud to announce that we have successfully released our EPAC1.0 Test Chip for fabrication.

Publication date: 1 June 2021

https://www.european-processor-initiative.eu/

- Vector-length agnostic ISA extensions (ARM-SVE, RISC-V Vector)

- CPUs with GPU-like capabilities and high energy efficiency



Supercomputer Fugaku CPU A64FX Realizing High Performance, High-Density Packaging, and Low Power Consumption

Ryohei Okazaki, Takekazu Tabata, Sota Sakashita, Kenichi Kitamura, Noriko Takagi, Hideki Sakata, Takeshi Ishibashi, Takeo Nakamura, Yuichiro Ajima

https://www.fujitsu.com/global/documents/about/resources/publications/technicalreview/2020-03/article03.pdf

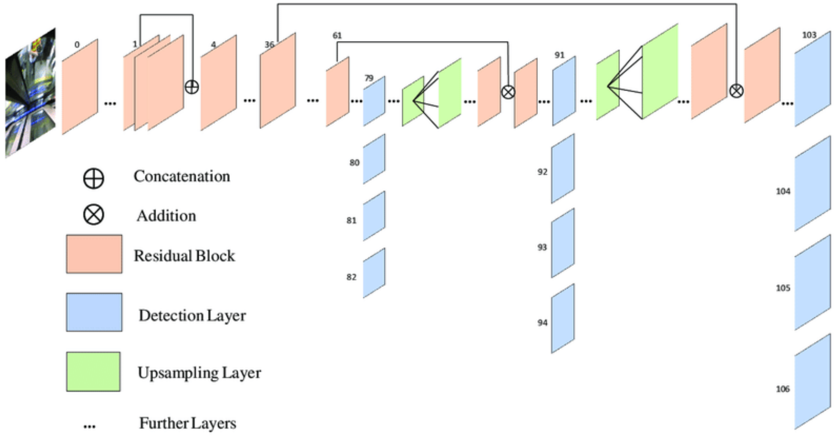
Matrix Multiplication on Vector Architectures

Case study: Convolutional Neural Networks

Multiple ways to implement convolutions:

- **im2col+GEMM**
- **Winograd**
- Direct
- FFT

GEMM appears in other layers as well



YOLOv3 object detection (inference):

>90% of total execution time spent on GEMM

[Sonia Rani Gupta et al., *Accelerating CNN inference on long vector architectures via co-design*, IPDPS 2023]

Matrix Multiplication on Vector Architectures

Can we optimize matrix multiplication for convolutions on emerging vector architectures?

- Algorithmic optimizations: Utilize the vector lengths and registers effectively

- Hardware parameters: Tune vector lengths, caches, and on-chip parallelism

Co-design convolutions and vector architectures for high-performance inference

[Sonia Rani Gupta et al., *Accelerating CNN inference on long vector architectures via co-design*, IPDPS 2023]

Matrix Multiplication on Vector Architectures

Algorithmic optimizations for im2col+GEMM

3-loops (baseline version)

- Vectorization using intrinsic instructions

- Multiple Multiply-add instructions

Maximize register utilization

Efficient reuse of vector registers

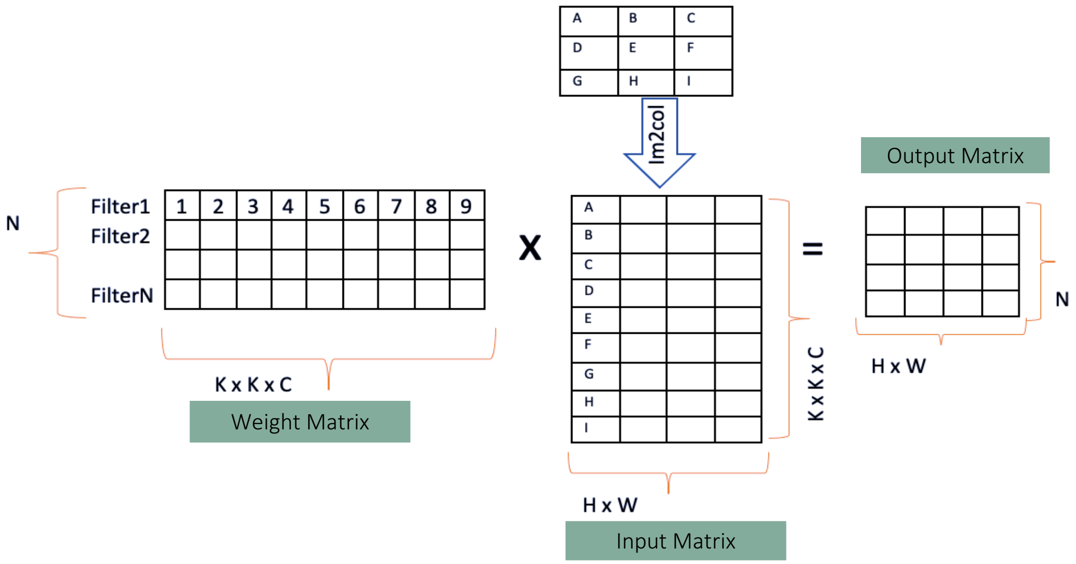
- Loop reorder

6-loops (Tiled version)

+ Packing of matrices

+ Tuning of block sizes block sizes

+ Prefetching



N: number of filters
K:kernel size
C: number of channels
H: height of image
W: weight of image

Impact of optimizations:

20-35x speedup on CNN inference on ARM A64FX

[Sonia Rani Gupta et al., *Accelerating CNN inference on long vector architectures via co-design*, IPDPS 2023]

Matrix Multiplication on Vector Architectures

Algorithmic optimizations for Winograd

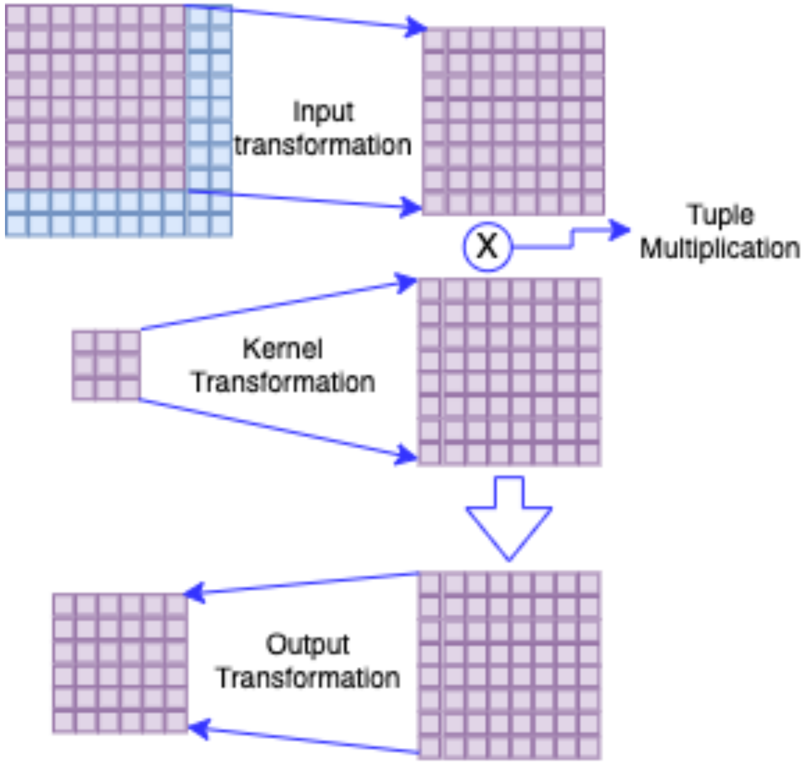
Transformations:

8x8 tile from one channel

inter-tile parallelism across the channels

Tuple multiplication:

Increase tuple size from 3 to 16
with 4 elements in each block
to utilize longer vector lengths



Impact of optimizations:

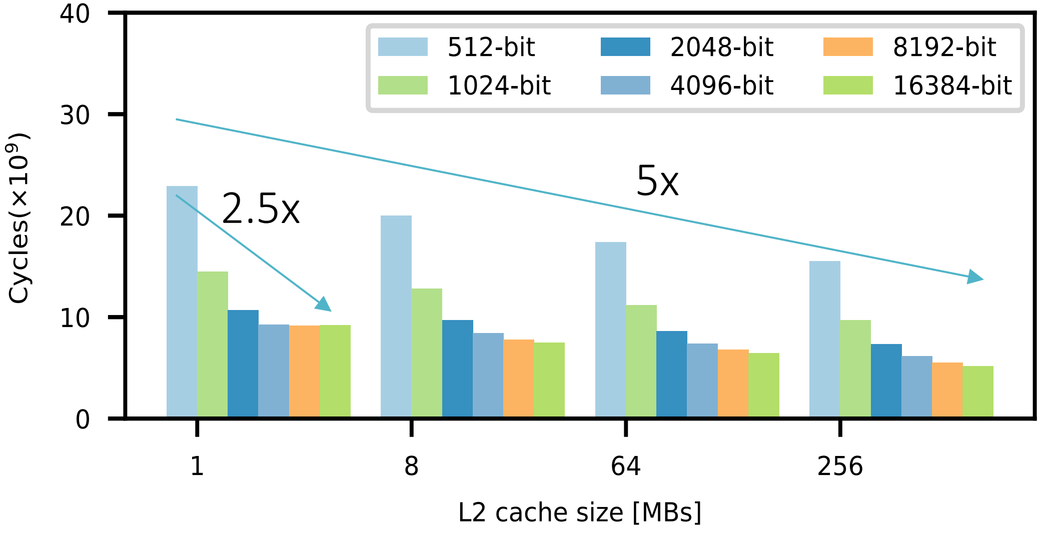
1.35-1.5x over im2col+GEMM

[Sonia Rani Gupta et al., *Accelerating CNN inference on long vector architectures via co-design*, IPDPS 2023]

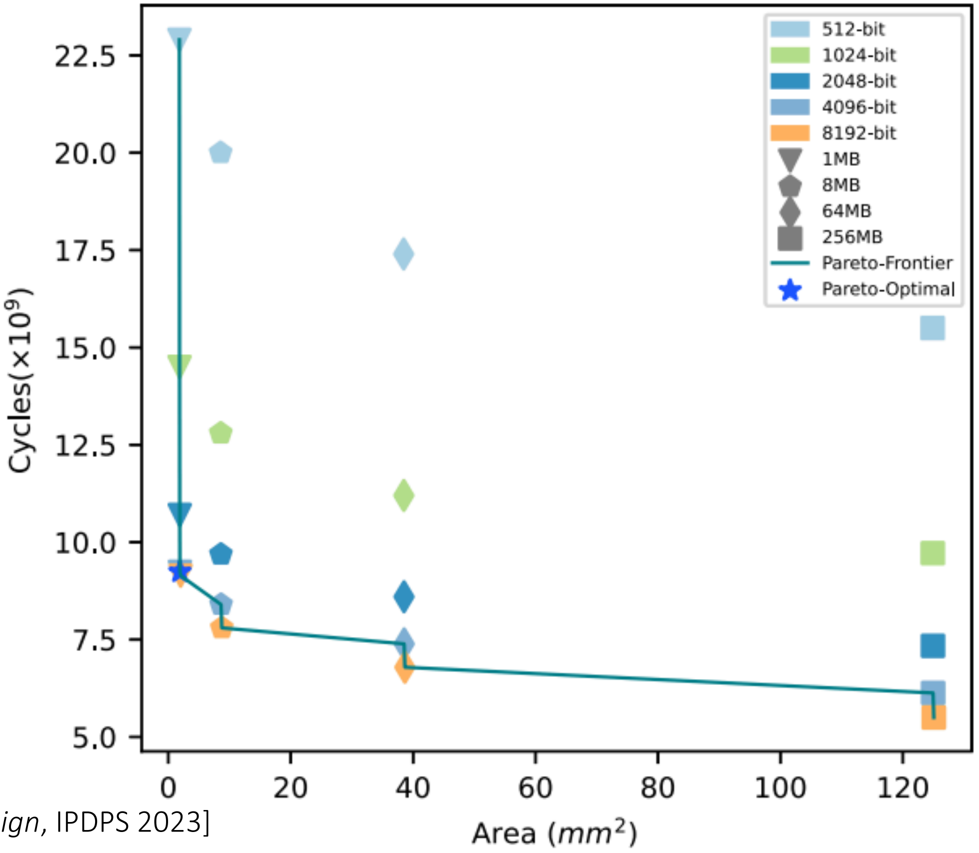
Matrix Multiplication on Vector Architectures

Hardware parameters and co-design

Impact of vector lengths on RISC-V Vector @ gem5 for YOLOv3 (20 layers) , for increasing L2 caches and 8 vector lanes.



Performance-area tradeoffs for a RISC-VV VPU with 8 lanes at 7nm FINFET



[Sonia Rani Gupta et al., *Accelerating CNN inference on long vector architectures via co-design*, IPDPS 2023]

Case #3

Inference serving

AI is high demand for inference

A trend that will increase because of GPT-like models

Inference runs as a service

Pre-trained network models on the cloud

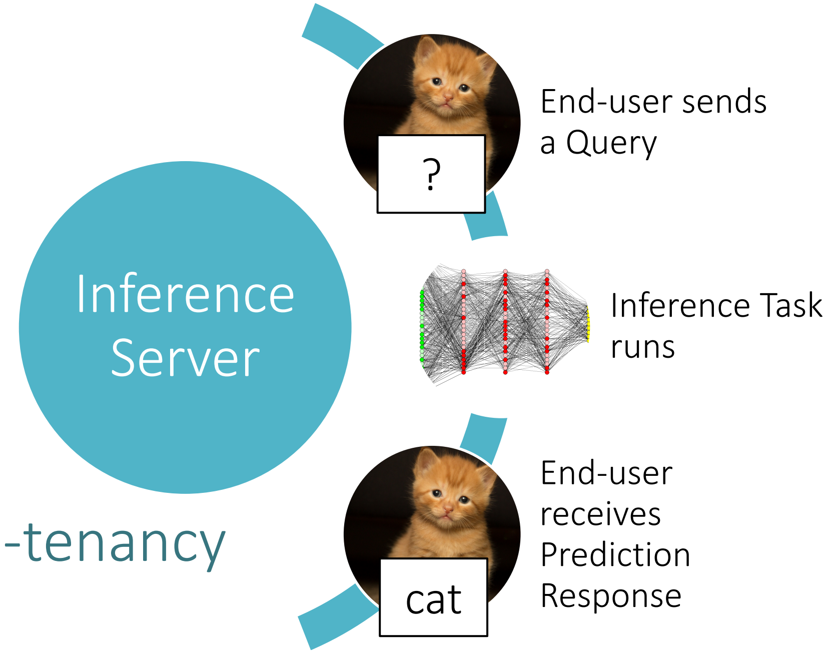
Independent inference queries from applications

Multiple users share the same inference server

Achieve high throughput and low latency despite multi-tenancy

Mitigate interference on inference pipelines

Assign resources dynamically



Inference Server

End-user sends a Query

Inference Task runs

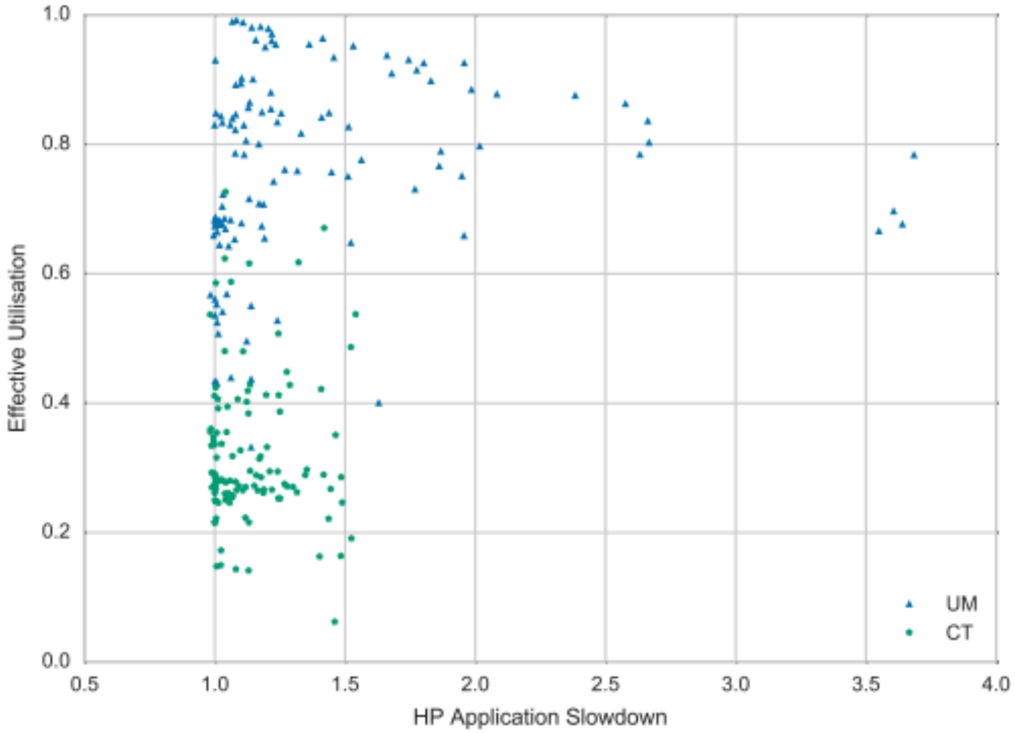
End-user receives Prediction Response

Inference serving

Can we consolidate more workloads alongside a high-priority inference service?

Challenges:

- Effectively partition resources
- Maintain high server utilization
- Maintain SLOs (latency)



[Konstantinos Nikas et al, *DICER: Diligent cache partitioning for efficient workload consolidation*, ICPP 2019]

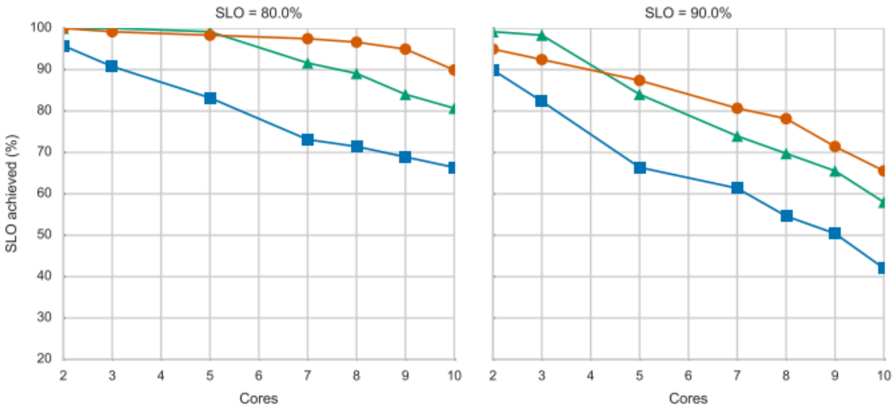
Inference serving

Managed resource allocation to favor the priority application

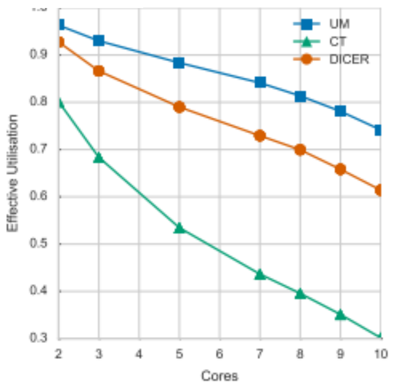
Boost the priority application without compromising the server utilization

1. Monitor consolidated applications

2. Appy dynamic cache partitioning
to mitigate cache and bandwidth contention



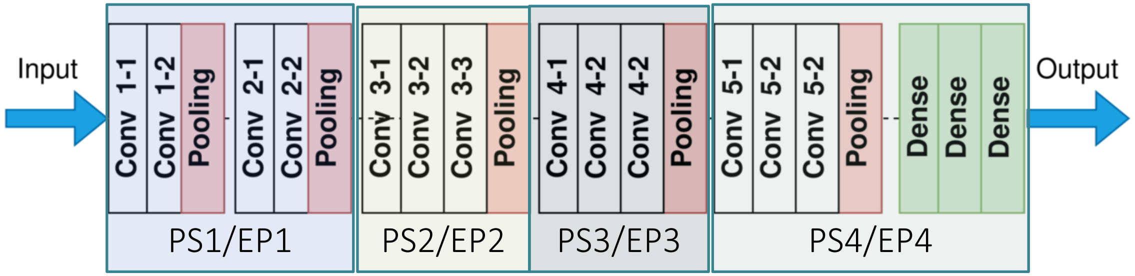
Improvements:
- 75% of workloads achieve SLOs of 90%
- Better throughput / server utilization with fewer violations

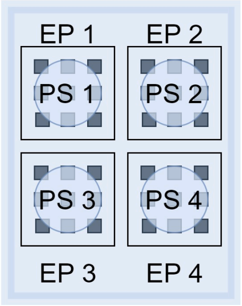


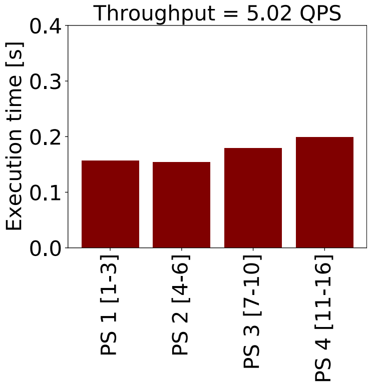
[Konstantinos Nikas et al, *DICER: Diligent cache partitioning for efficient workload consolidation*, ICPP 2019]

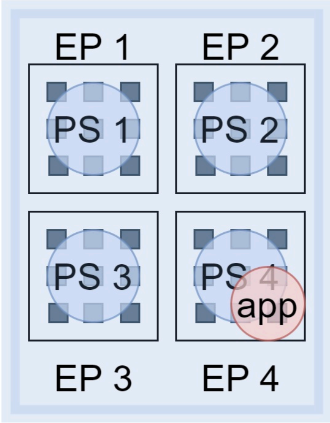
Inference serving

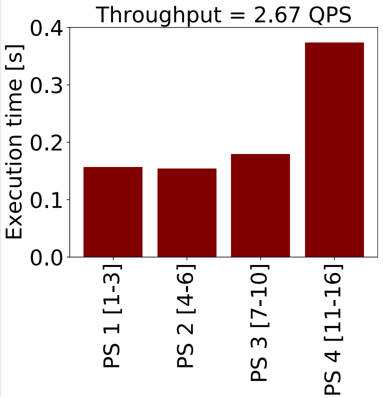
Can we adapt the resources of the inference pipeline to avoid interference?

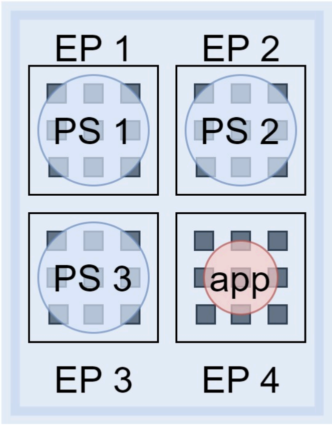


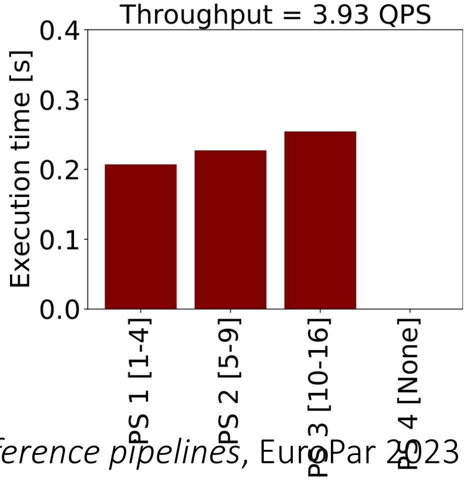


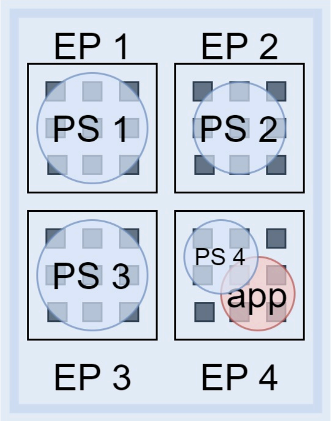


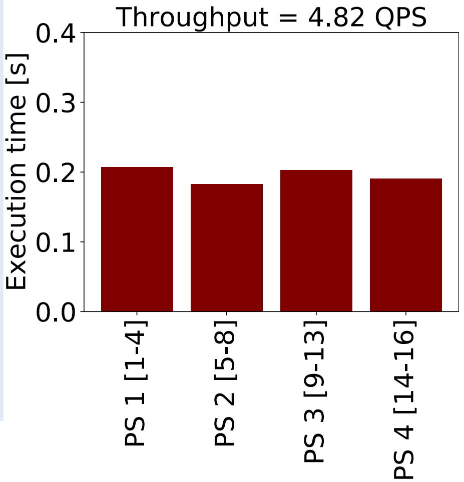












[Pirah Noor Soomro et al., *ODIN: Overcoming Dynamic Interference in iNference pipelines*, EuroPar 2023]

Inference serving

Reactively balanced inference pipelines that avoid interference

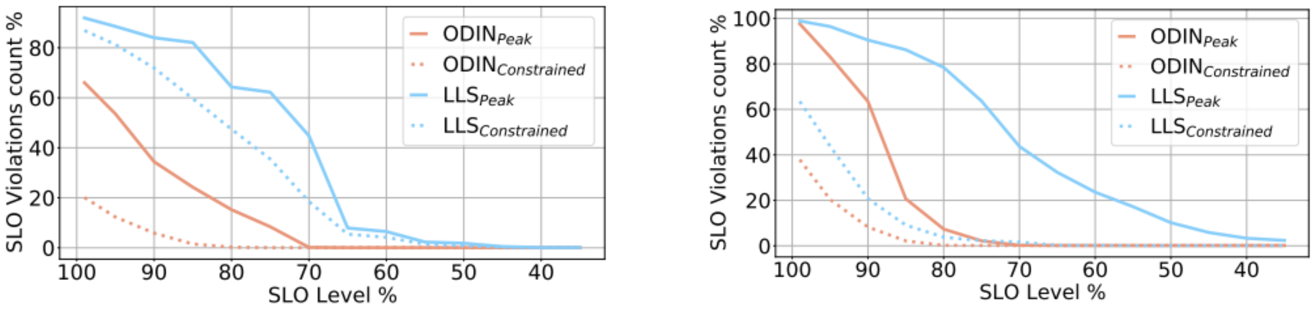
Maximize throughput despite interference from co-located applications

1. Target balanced parallel inference pipelines that offer high throughput
2. Dynamically detect interference
3. Rebalance the pipeline through heuristics

Improvements:

- 15% better latency and 20% better throughput compared to least-loaded scheduler

- Avoid 80% of SLO violations at an 80% SLO level



(a) ResNet50

(b) VGG16

[Pirah Noor Soomro et al., *ODIN: Overcoming Dynamic Interference in iNference pipelines*, EuroPar 2023]

Case #4

The future is sparse

The future is sparse (in resources)

AI models are growing rapidly

  Billions of parameters

PFLOPs (soon ExaFLOPs) to train

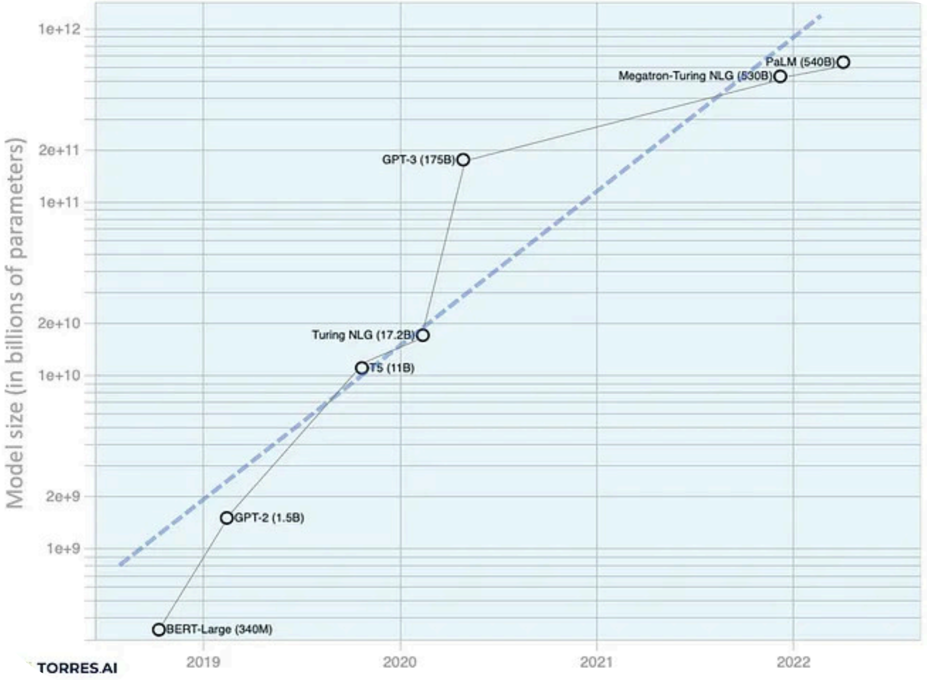
  Major companies investing in supercomputers

Inference will cost more

  High demand in applications

  Already 80-90% of AWS and NVIDIA cloud

Environmental impact is currently unknown!



https://towardsdatascience.com/transformers-the-bigger-the-better-19f39f222ee3

AI future needs to be sparse

Neural network computations can be sparsified

- Reduce connections between neurons
- Orthogonal to quantization

Performance Impact

- Tolerable sparsity: 50-95% [1]
  - ~**20x** reduction in FLOPs
  - ~**10x** potential performance improvement

[1] Hoefler, Torsten, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. "Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks." *The Journal of Machine Learning Research* 22, no. 1 (2021): 10882-11005.

Sparse Matrix-Vector Multiplication

Feature-based performance analysis and characterization

Understand characteristics of sparse matrices and modern devices that determine performance and energy efficiency

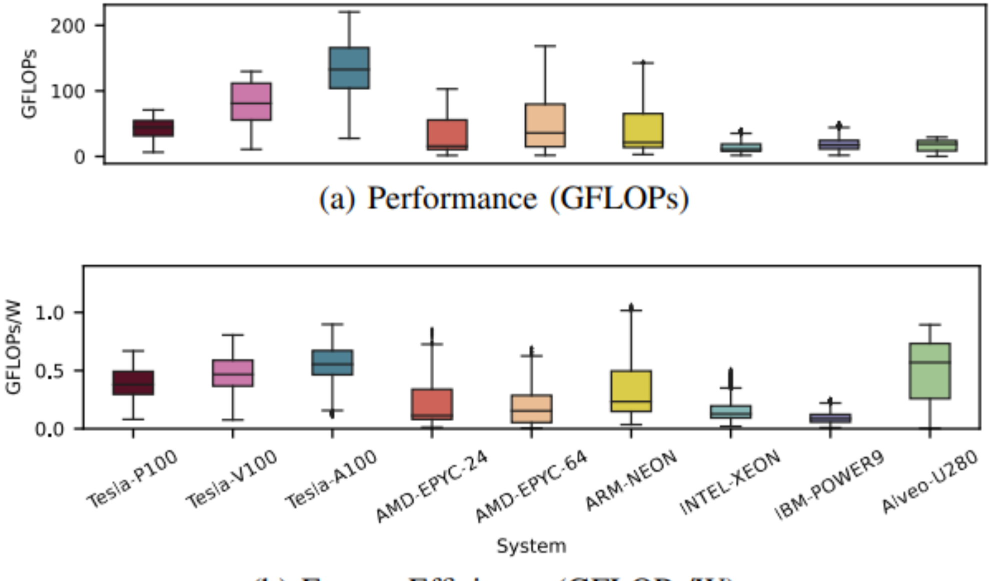
Artificial sparse matrix generator

Sparsity features associated with performance

Takeaways:

- GPUs prevail, but CPUs are competitive for smal matrices, FPGAs are competitive in energy

- SpMV is memory-bound but low ILP and low operational intensity are of concern



(a) Performance (GFLOPs)

(b) Energy Efficiency (GFLOPs/W)

[Panagiotis Mpakos et al, *Feature-based SpMV performance analysis on contemporary devices*, IPDPS 2023]

Making the AI future sustainable

2030 AI should be SpMM: sparse matrix-matrix multiplication

Obstacle #1: Modern architectures are not optimized for sparsity

Hardware-aware optimizations!

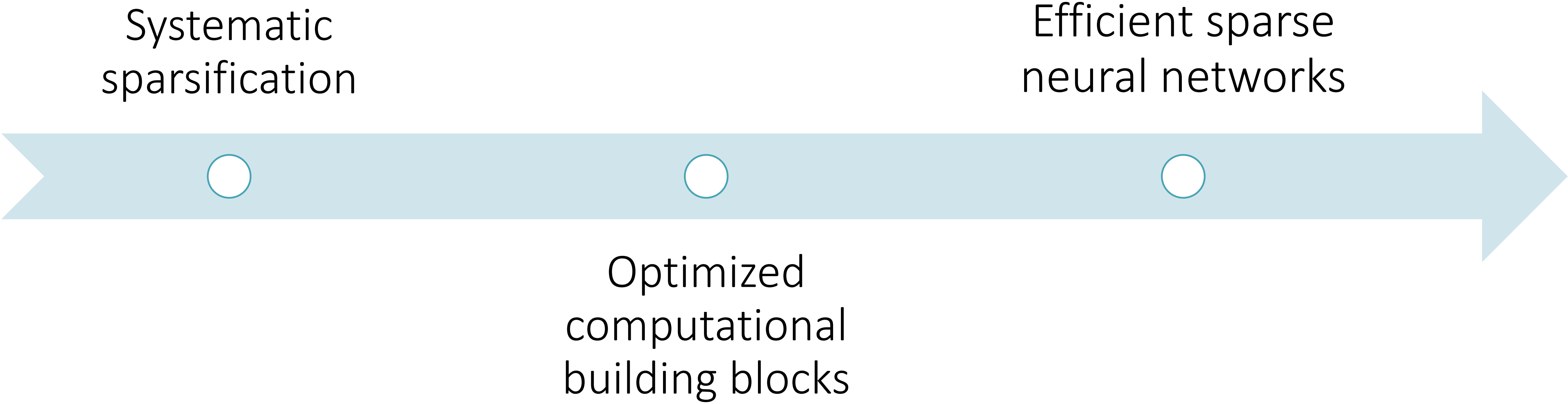
Obstacle #2: Sparsity in neural networks is different from HPC

Autotuning!

Making the AI future sustainable

Follow the “algorithmic” Moore’s law – don’t leave performance on the table

Hardware will not improve fast enough, but software can close some of the gaps

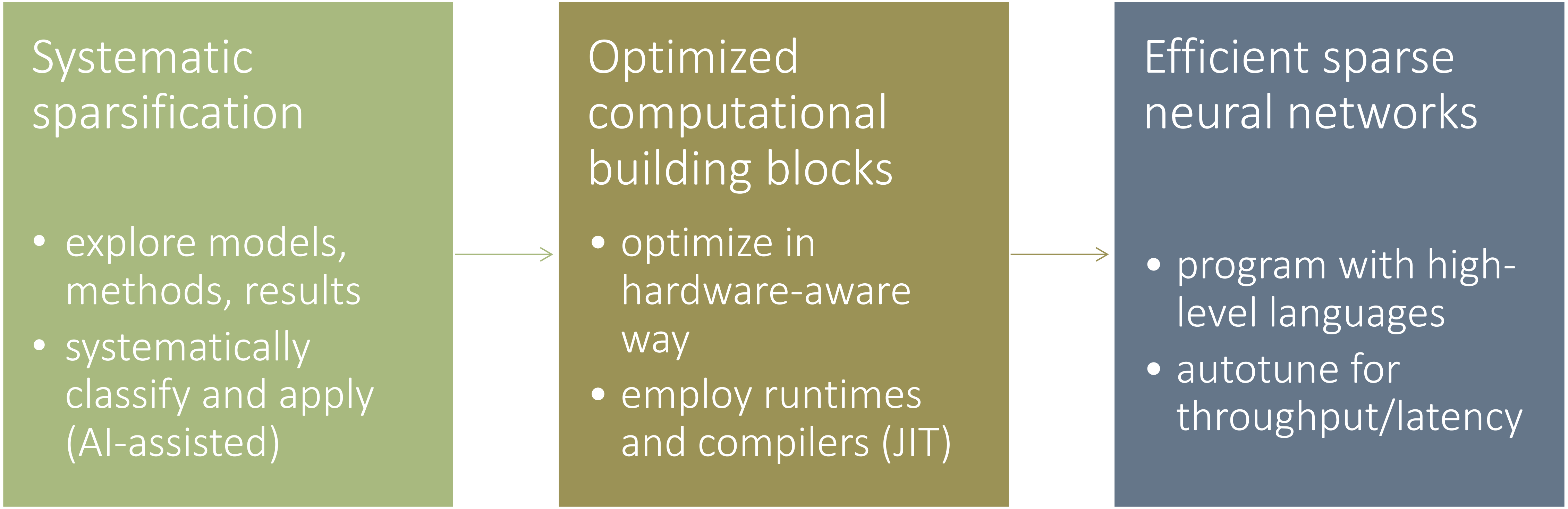


Systematic sparsification

Optimized computational building blocks

Efficient sparse neural networks

Sustainable AI: a research proposal



Systematic sparsification

- explore models, methods, results
- systematically classify and apply (AI-assisted)

Optimized computational building blocks

- optimize in hardware-aware way
- employ runtimes and compilers (JIT)

Efficient sparse neural networks

- program with high-level languages
- autotune for throughput/latency

Sustainable AI: impact

- Potential performance improvements of 10x
- Potential energy savings of 10x in training and inference
- Reduced need for high-end power-hungry hardware
- CO2 emission reduction
- Broader access for users/industry

Publications / Acknowledgements

Petros Anastasiadis, Nikela Papadopoulou, Georgios I. Goumas, Nectarios Koziris. **CoCoPeLia: Communication-Computation Overlap Prediction for Efficient Linear Algebra on GPUs**, *IEEE International Symposium on Performance Analysis of Systems and Software, ISPASS 2021, Stony Brook, NY, USA, March 28-30, 2021*, 2021, DOI: *https://doi.org/10.1109/ISPASS51385.2021.00015*

Petros Anastasiadis, Nikela Papadopoulou, Georgios Goumas, Nectarios Koziris, Dennis Hoppe, Li Zhong. **PARALiA: A Performance Aware Runtime for Auto-tuning Linear Algebra on heterogeneous systems**, *ACM Transactions on Architecture and Code Optimization*, 2023, DOI: *https://doi.org/10.1145/3624569*

Sonia Rani Gupta, Nikela Papadopoulou, Miquel Pericàs. **Accelerating CNN inference on long vector architectures via co-design**, *IEEE International Parallel and Distributed Processing Symposium, IPDPS 2023, St. Petersburg, FL, USA, May 15-19, 2023*, 2023, DOI: *https://doi.org/10.1109/IPDPS54959.2023.00024*

Sonia Rani Gupta, Nikela Papadopoulou, Miquel Pericàs. **Challenges and Opportunities in the Co-Design of Convolutions and RISC-V Vector Processors**, *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, 2023, DOI: *https://doi.org/10.1145/3624062.3624232*

Konstantinos Nikas, Nikela Papadopoulou, Dimitra Giantsidi, Vasileios Karakostas, Georgios I. Goumas, Nectarios Koziris. **DICER: Diligent Cache Partitioning for Efficient Workload Consolidation**, *Proceedings of the 48th International Conference on Parallel Processing, ICPP 2019, Kyoto, Japan, August 05-08, 2019*, 2019, DOI: *https://doi.org/10.1145/3337821.3337891*

Pirah Noor Soomro, Nikela Papadopoulou, Miquel Pericàs. **ODIN: Overcoming Dynamic Interference in iNference Pipelines**, *Euro-Par 2023: Parallel Processing - 29th International Conference on Parallel and Distributed Computing, Limassol, Cyprus, August 28 - September 1, 2023, Proceedings*, 2023, DOI: *https://doi.org/10.1007/978-3-031-39698-4_12*

Panagiotis Mpakos, Dimitrios Galanopoulos, Petros Anastasiadis, Nikela Papadopoulou, Nectarios Koziris, Georgios I. Goumas. **Feature-based SpMV Performance Analysis on Contemporary Devices**, *IEEE International Parallel and Distributed Processing Symposium, IPDPS 2023, St. Petersburg, FL, USA, May 15-19, 2023*, DOI: *https://doi.org/10.1109/IPDPS54959.2023.00072*

Thank you!

Questions?