MAGUS: Memory-Throughput-Based Uncore Frequency Scaling on Heterogeneous Systems

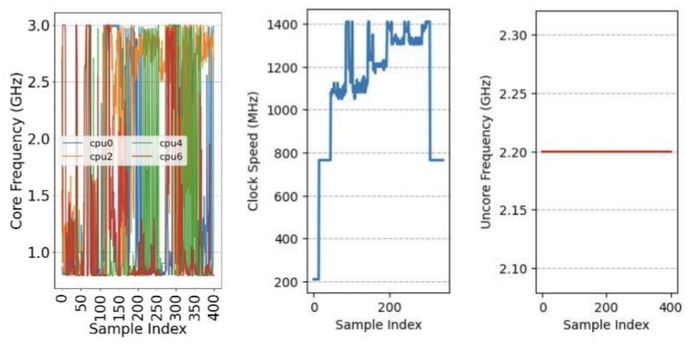
Zhong Zheng, Seyfal Sultanov, Michael E. Papka, Zhiling Lan

Motivation & Illustrative Example

Prior uncore frequency tuning studies have primarily focused on conventional HPC workloads running on CPU-only systems. As HPC advances toward heterogeneous computing, integrating diverse GPU workloads on heterogeneous CPU-GPU systems, it becomes imperative to revisit and enhance uncore scaling. Our investigation reveals that uncore frequency scales down only when CPU power approaches its TDP (Thermal Design Power) --- a rare scenario in GPU-dominant applications --- resulting in unnecessary power waste in modern computing systems.

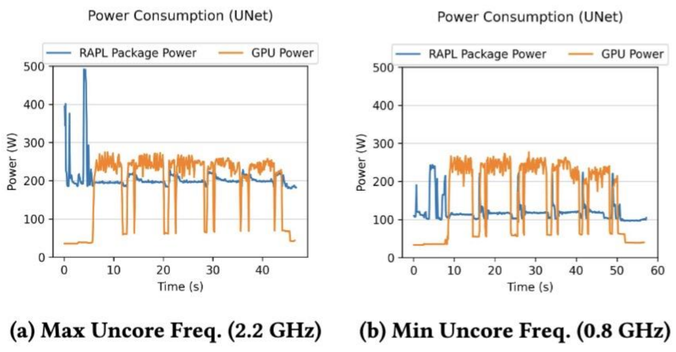
UNet Training Example

- A heterogeneous system with Intel Xeon CPU–A100 GPU node
- CPU core frequency and GPU clock speed are dynamically adjusted by default
- Uncore frequency remains at its maximum



(a) CPU freq. (b) GPU clock speed (c) Uncore freq.

- Power profiles of UNet training under different uncore frequencies: max (2.2 GHz) versus min (0.8 GHz).
- Reducing the uncore frequency results in (i) an 82-watt reduction in CPU power, from 200 watts (blue curve on the left) to 120 watts (blue curve on the right).
- An increase in runtime, from 47 seconds (left) to 57 seconds (right).



(a) Max Uncore Freq. (2.2 GHz) (b) Min Uncore Freq. (0.8 GHz)

Experimental Setup

Heterogeneous systems

- **Intel+A100**: A Chameleon Cloud [37] system featuring two Intel(R) Xeon(R) Platinum 8380 processors paired with a single NVIDIA A100-40GB GPU.
- **Intel+4A100**: It has the same architecture and software environment as the first, except it is equipped with 4x NVIDIA A100-80GB GPUs interconnected via PCIe.
- **Intel+Max1550**: It features the Intel(R) Xeon(R) CPU Max 9462, a Sapphire Rapids architecture processor comprising 8x compute tiles Intel(R) Data Center GPU Max 1550 based on the Ponte Vecchio architecture.

Benchmarks & Applications

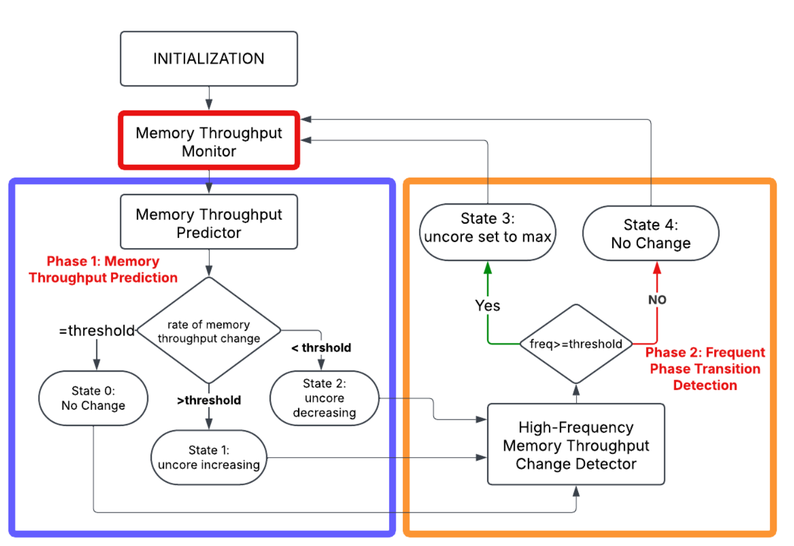
- **GPU benchmark suite Altis:** Level 1 & Level2 applications
- **ECP Proxy Apps**: miniGAN, CRADL, Laghos, SW4lite
- **AI-enabled applications:** GROMACS, LAMMPS
- **MLPerf training Benchmarks:** UNet, Bert, Resnet50

Evaluation Metrics

- **Performance loss**: percentage increase in execution time compared to the baseline
- **Power saving:** CPU package + DRAM power saving
- **Energy saving:** CPU package + DRAM + GPU energy saving

MAGUS Design

- Phase 1: Memory Throughput Prediction
- Phase 2: Frequent Phase Change Detection



Algorithm 1: Memory Throughput Trend Prediction

```
function MEM_THT_TREND_PREDICTION(direv_length, inc_threshold, dec_threshold, mem_throughput_ls)
    derivative ← (mem_throughput_ls[−1] − mem_throughput_ls[0]) / direv_length
    if derivative > inc_threshold then
        return 1
    else if derivative < dec_threshold then
        return -1
    else
        return 0
    end if
end function
```

Algorithm 2: High Frequency Detection

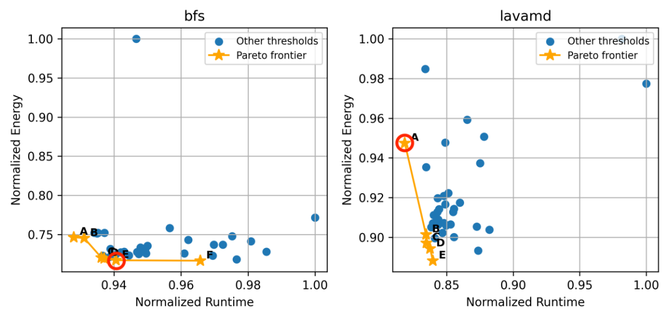
```
function HIGH_FREQ_DETECTION(high_freq_threshold, uncore_tune_ls)
    freq ← sum(uncore_tune_ls) / len(uncore_tune_ls)
    if freq ≥ high_freq_threshold then
        return True
    else
        return False
    end if
end function
```

Overheads

System	Power Overhead (%)		Invocation Overhead (s)	
	MAGUS	UPS	MAGUS	UPS
Intel + A100	1.1%	4.9%	0.1s	0.3s
Intel + Max1550	1.16%	7.9%	0.1s	0.31s

Sensitivity Analysis

- Pareto frontiers of energy consumption and runtime under different threshold configurations.
- The red circled Pareto frontier represents the common threshold set observed across all applications tested in our experiments.

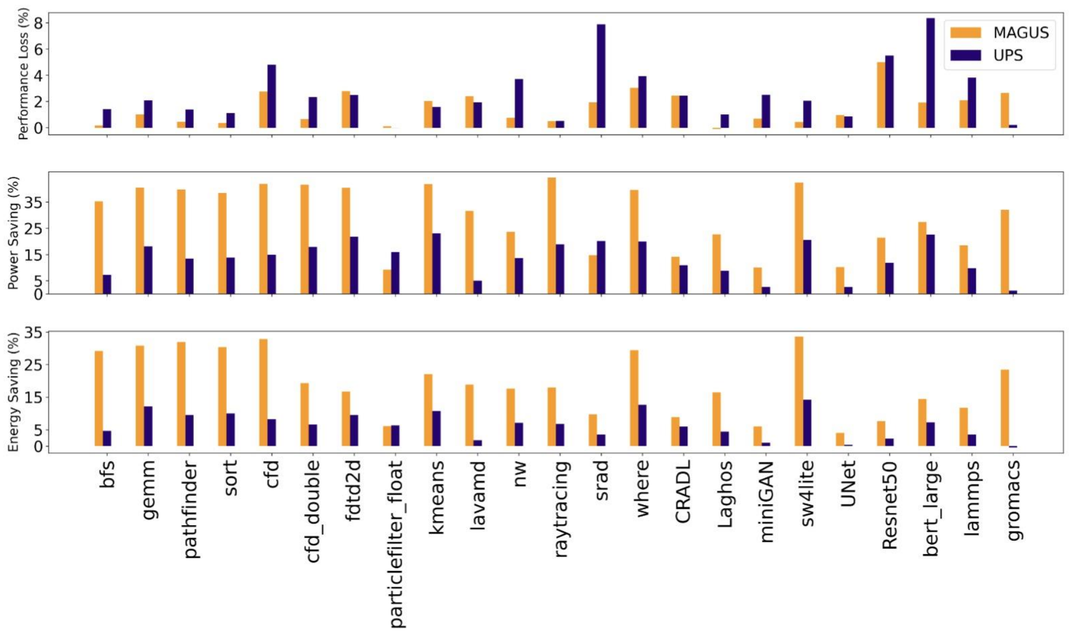


Prediction Accuracy

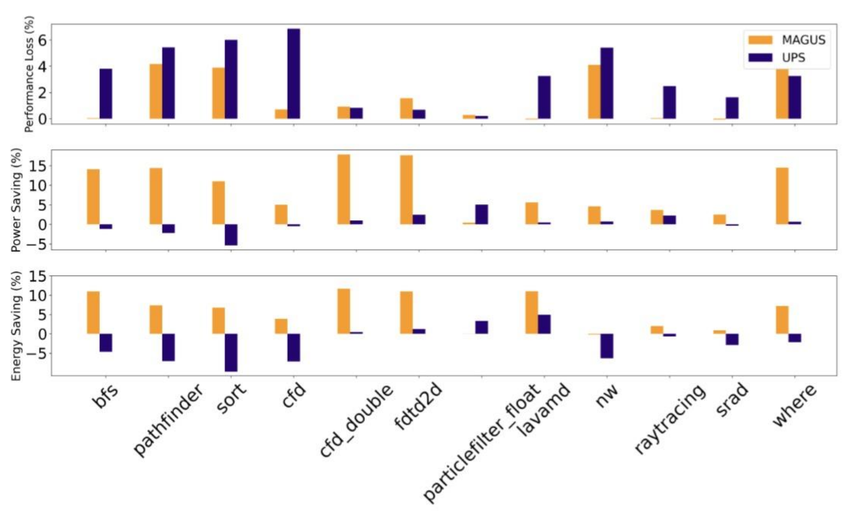
Table 1: Jaccard similarity for memory throughput trend

Application	Jaccard	Application	Jaccard
bfs	0.99	gemm	0.71
pathfinder	0.98	sort	0.96
cfd	0.94	cfd_double	0.63
fdtd2d	0.40	kmeans	0.97
lavamd	0.92	nw	0.98
particlefilter_float	0.67	raytracing	0.87
where	0.94	Laghos	0.99
miniGAN	0.98	sw4lite	0.87
UNet	0.99	Resnet50	0.96
bert_large	0.84	lammps	0.99
gromacs	0.99		

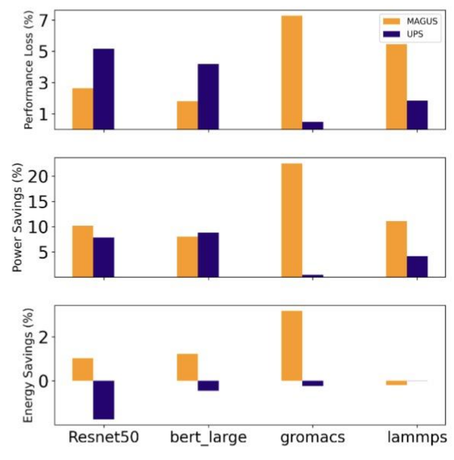
End-to-End Performance



(a) Overall performance on Intel+A100.



(b) Overall performance on Intel+MAX1550.



(c) Overall performance on Intel+4A100.

UNIVERSITY OF ILLINOIS CHICAGO

Argonne NATIONAL LABORATORY