HYPER ACCEL

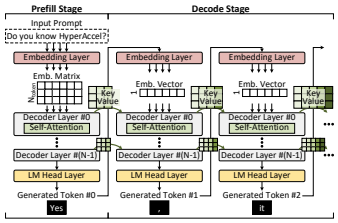
Adelia: A 4nm LLM Processor for Efficient Generative AI Inference

Seungjae Moon, Jung-Hoon Kim, Juntaek Oh, Jay Kim, Joo-Young Kim

HyperAccel, Seoul, South Korea

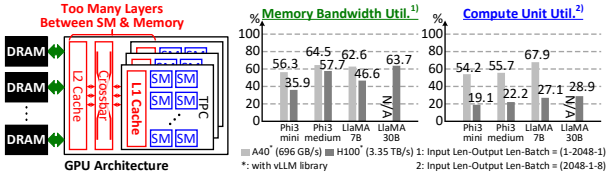
Background & Motivation

LLM Structure



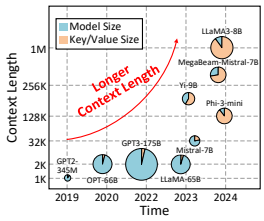
1. Computational imbalance between LLM stages
   - Prefill = GEMM = compute bottleneck
   - Decode = GEMV = memory bottleneck
   - KV cache computed in prefill then continuously appended in decode

GPU Shortcomings



2. Low hardware utilization of GPU
   - Dataflow with complex hierarchy for programmability
   - Consistent data load/store due to little core
   - Massively parallel cores optimized for training, not inference
   - Unnecessary power consumption and high device cost

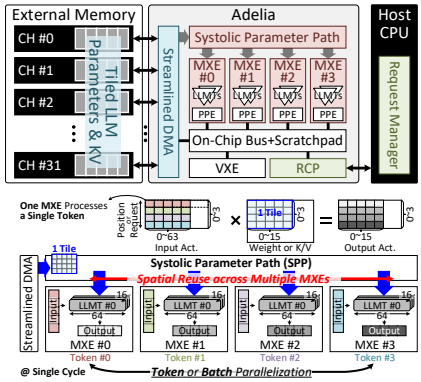
LLM Trend



3. Growth of model parameters and KV cache
   - Increase in computational intensity
   - Demand for scalable hardware
   - Quantization to reduce memory

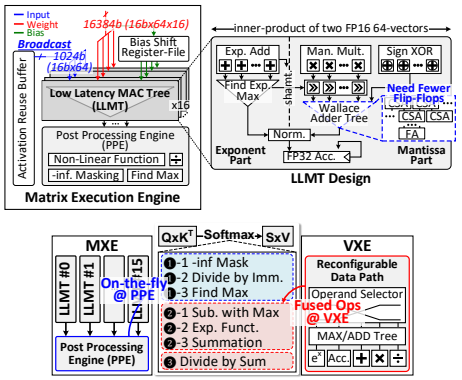
Adelia Architecture

Overall Architecture



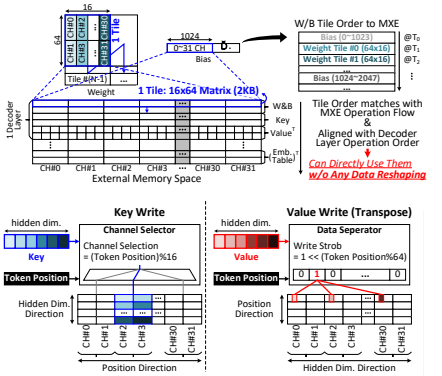
1. Streamlined dataflow for maximum hardware utilization
   - Perfect alignment of memory and compute bandwidth to remove unnecessary buffers
   - Systolic path for spatial reuse of parameters and KV
   - Reduced data load/store due to large core
   - Fully scalable hardware to leverage growing LLM

Microarchitecture



2. LLM-optimized features
   - MAC with hardware sharing to support multi-precision, FP16, BF16, FP8, INT8, INT4, at reduced power and area
   - Fused operations supported by additional processors to hide latency of bottleneck operations

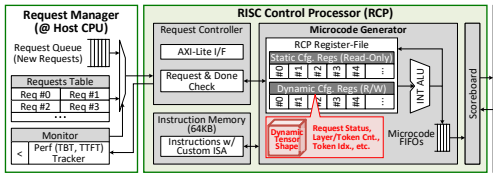
Software-Hardware Co-Design



3. Data preprocessing
   - Hardware-aware mapping of parameters for burst read of data from memory without data reshaping
   - Use of write strobe to handle dynamic movement of activations without physical or performance overhead

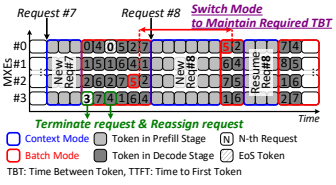
Scheduler

Controller



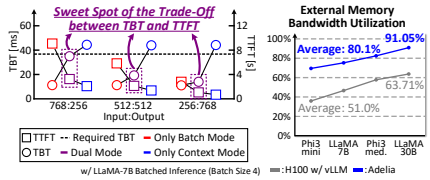
1. Runtime management of user request
   - Continuous monitoring of request information to issue instructions and configuration signals for high efficiency
   - Flexible broadcast of input based on performance requirement

Dual-Mode Parallelization



2. Load balancing based on service-level objective
   - Context mode: prioritize single user for low latency
   - Batch mode: prioritize multi-user for high throughput
   - Support for continuous batching

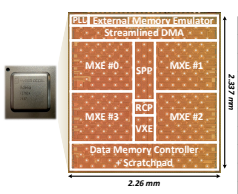
Performance Improvement



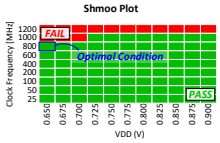
3. Achievement of optimal performance and efficiency
   - Ability to configure hardware to meet required TBT
   - Unprecedented memory bandwidth utilization of ~90%

Evaluation

Adelia Specification



Shmoo Plot



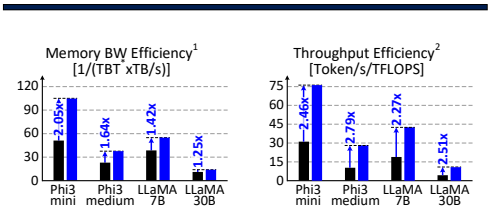
Technology		4 nm SF4X FinFET
Die Area		5.28 mm²
On-Chip Memory		2.16 MB
Supply Voltage		0.65 V ~ 0.9 V
Clock Frequency		25 MHz ~ 1 GHz
Data Precision		FP16
[1) 3)] Peak Throughput		8.19 TFLOPS
[2) 3) 4)] Peak Energy Efficiency		25.3 TFLOPS/W
[1) 3)] Peak Area Efficiency		1.55 TFLOPS/mm²
Target Model		Transformer (LLM)
Parallelism		Batch + Context
[1) 5) 6)] Inference Throughput [Token/s]	Phi3-mini	512.60
	Phi3-medium	184.33
	LLaMA-7B	258.21
	LLaMA-30B	65.72

1) @ 1GHz, 0.725V
2) @ 800MHz, 0.65V
3) 1MAC = 2Operations
4) EMA is excluded
5) EMA is included (estimated with HBM2E, 1.64TB/s)
6) Input length: 256, output length: 768, 4 Batches

Inference Throughput

Model	Batch Size	Token Length [In-Out]	Runtime [s]	Inference Throughput [Token/s]
Phi3-mini	1	768-256	2.397	106.79
	2	512-512	4.263	240.17
	4	256-768	5.993	**512.60**
Phi3-medium	1	768-256	6.902	37.09
	2	512-512	12.099	84.63
	4	256-768	16.666	184.33
LLaMA 7B	1	768-256	4.930	51.93
	2	512-512	8.641	118.51
	4	256-768	11.897	258.21
LLaMA 30B	1	768-256	19.778	12.94
	2	512-512	34.367	29.80
	4	256-768	46.741	**65.72**

Efficiency



- At HyperAccel, Adelia is available as a hardware IP that enables LLM in any SoC due to its ultimate scalability.
- Adelia achieves up to x higher normalized throughput than NVIDIA H100.
- Adelia achieves up to x less normalized power consumption than NVIDIA H100.
- With additional improvements, our upcoming singe-run, Bertha, will be released as an accelerator card with unprecedented efficiency to reduce the TCO of datacenters when running LLM inference.