SPEC SHEET

IMEC.KELIS

This section describes imec.kelis as delivered before 2020904 and how it was planned to evolve to 1.0.0. A beta version is published in the imec cloud. An alpha (0.1.4) and beta (0.3.0) version are available as a container.

Top level

Imec.kelis (formerly (Optimus') is an analytical performance modeling tool for large language models running on distributed compute systems. Visit www.imec-int.com/kelis for more info.

The user interacts with imec.kelis using a web interface. In version 1.0.0, an API will be exposed that allows scripts to call interfaces inside imec.kelis directly without interaction with the web interface. In version 1.0.0 the available network topologies will have been extended.

Configurability

The user describes the system to be analyzed using a predefined set of parameters for the application or workload and system architecture. The user can store several configurations, execute the analysis and compare the results.

Application or workload

For the workload, imec.kelis offers configurable models for decoder-based bransformers, such as GPT and llamas. Both the training and inference can be modeled.

System composition

For composing a system, imec.kelis offers the following configurable components:

COMPONENT	AVAILABLE COMPONENTS	VALIDATED
GPU	Nvidia V100, B200, H200, H100, A100, AMD MI300X (experimental)	V100, A100, H100
CPU	AMD epyc, intel xeon platinum 8480C	N.A.
DRAM	DDR4, DDR5, HBM2(e), HBM3(e), HBM4	HBM2(e), 3
NIC	Infiniband, Mellanox ConnectX6, ConnectX7, ConnectX8	Infiniband, MC6
Network Topologies (Experimental)	fattree, mesh, torus, ring, switch-based	Y
PCIe switch	gen4, gen5	N.A.
Scale up network switch	NVlink Switch gen2, gen3	NVSwitch 2
SSD	2 density models	N.A.

SPEC SHEET

IMEC.KELIS

Workload

1. The workload parallelization is modeled as multidimensional parallelization mapper based on an LLM task graph including DP, TP, PP.
2. The LLM task graph consists of six levels of GEMM. Four levels are dedicated to self-attention and two are dedicated to MLP.
3. For the workload modeling, a hierarchical roofline for tiled GEMM/V is used. The model includes KV caching.
    - The GEMM sizes depend on the parallelism choice.
    - For every GEMM, tiling is performed at every memory level and the optimal tiling is selected. Computation time is calculated from the roofline model.
4. Performance models for collective communication algorithms used during distributed LLM deployments are available.

Performance reporting

The following results are shown in charts or tables:

- Individual workload execution time breakdown
    - Communication versus compute
    - Execution time due to parallelization methods
- Individual workload memory usage breakdown
- Kernels are plotted on roofline chart
- Multiple workloads over a system lifetime

Available LLMs

TYPE	NAME
GPT	GPT2, GPT3
LLama	Llama2, Llama3
MoE	Llama4-scout, Llama4-Maverick

Node scaling (experimental)

This feature enables users to explore the impact of different logic nodes and memory nodes.

We apply area and power scaling factors to derive the performance of different nodes. The factors are obtained from imec's fabs.

Logic scaling from N14 to N2.
HBM scaling from HBM2 to HBM3e.