Kung Flopanda 6070: The Next Generation of Supercomputing

Chief Engineers: Bryce Chen & Henry Jung

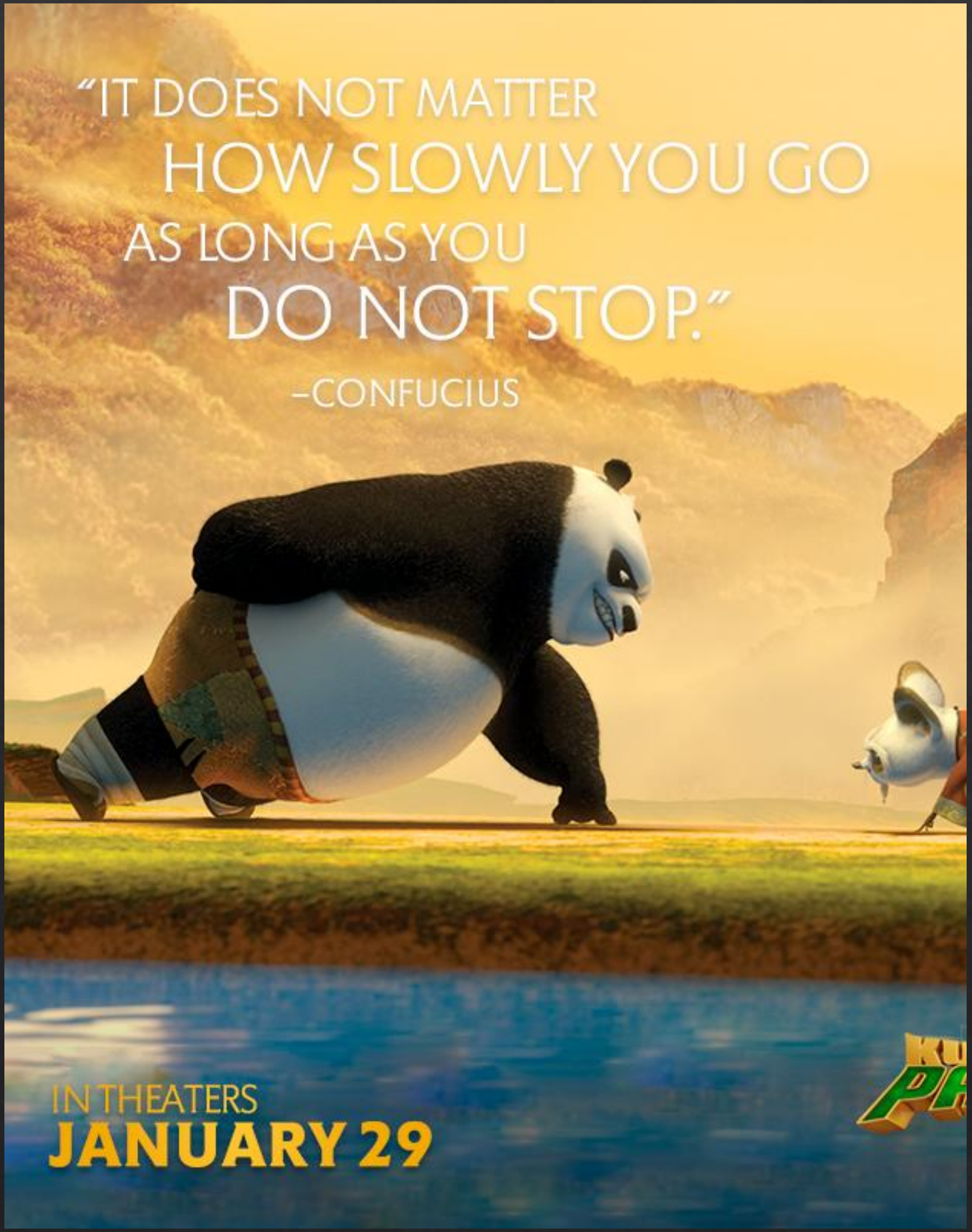
Executive Summary

- Kung FLOPanda (KFP LLC) is a company focused on architecture research for next-generation high-performance computing (HPC) through an analysis of current state of the art
- KFP has received a request for proposal for an HPC system for a release in 2030.

Master Oogway

Dragon Warrior

The Tigress

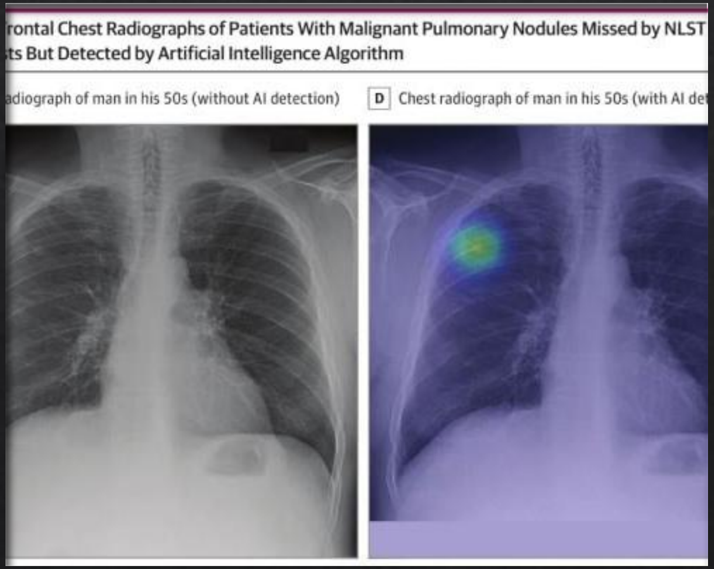


Motivation

- Increasing AI workload demands
- Growing data and model sizes
- Protecting the environment
- Saving the otters

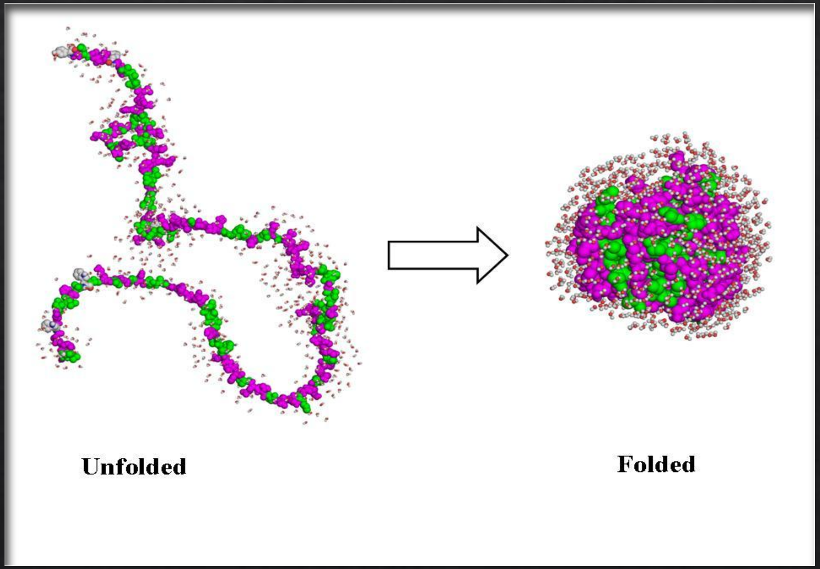
Impacts

- Accelerated scientific discovery
- AI model advancement
- Better medical diagnosis
- Efficient systems
- Improvement in climate prediction


rontal Chest Radiographs of Patients With Malignant Pulmonary Nodules Missed by NLST
ts But Detected by Artificial Intelligence Algorithm

adiograph of man in his 50s (without AI detection)

D Chest radiograph of man in his 50s (with AI det



Requirements

RocBLAS	OLLaMa	nBody	System Power	System Cost
Minimum 9.0 ExaFLOPs /sec	Minimum 3.2 Million Tokens /sec	Minimum 1.5 ExaFLOPs /sec	UNDER 25 MegaWatts	UNDER $550 Million

RocBLAS: LINPACK Equivalent (Linear Algebra and Matrix Operations)

OLLaMA: Large Language Model and Generative AI

nBody: Dynamical system of particles

Assumptions

RocBLAS	OLLaMa	nBody	Node Srink	Software Optimization
100% scaling factor	88% scaling factor	90% scaling factor	30% power saving & 15% frequency improvement	40% improvement

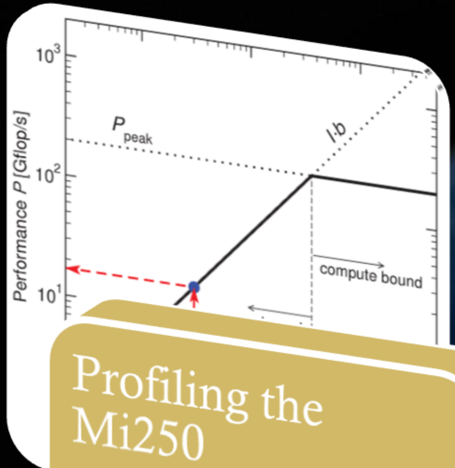
NODE SHRINKS: 2 SHRINKS

MEMORY BANDWIDTH: DOUBLING LEADS TO 20% POWER AND 15% COST INCREASE

VALU: DOUBLING LEADS TO 40% POWER AND 20% COST INCREASE

MFMA: DOUBLING LEADS TO 40% POWER AND 25% COST INCREASE

Methodology



Profiling the Mi250

- Roofline Analysis
  - Compute bound & Memory Bound
- Benchmark Analysis
  - Compute to Memory Ratio

Mathematical Modeling

- “Tweaking” the performance knobs
- Calculate improvements and trade offs

Find an architecture

- Meet the requirements
- Optimize based on the target
- Repeat for multiple solutions

Master Oogway

“Efficiency is true dominance”

Architecture Details

RocBLAS (ExaFLOPs)	OLLaMa (Tokens/Sec)	nBody (ExaFLOPs)	FP32 MFMA (TeraFLOPs)	FP32 VALU (TeraFLOPs)	HBM Bandwidth (TB/s)
9.01	3.30 Million	1.50	3868.65	831.977	610.18

MFMA	VALU	HBM
60x	52x	63x

# of Units	Unit POWER (W)	Total Power (W)
1181	32,144	18,600,000

The Dragon Warrior

"Balance is true strength"

Architecture Details (KFP 6070 Q)

RocBLAS (ExaFLOPs)	OLLaMa (Tokens/Sec)	nBody (ExaFLOPs)	FP32 MFMA (TeraFLOPs)	FP32 VALU (TeraFLOPs)	HBM Bandwidth (TB/s)
10.83	3.32 Million	2.06	4513	1007.9	242.14

MFMA	VALU	HBM
70x	63x	25x

# of Units	Unit POWER (W)	Total Power (W)
1337	32,592	21,090,000

The Tigress

"Precision driven MFMA dominance"

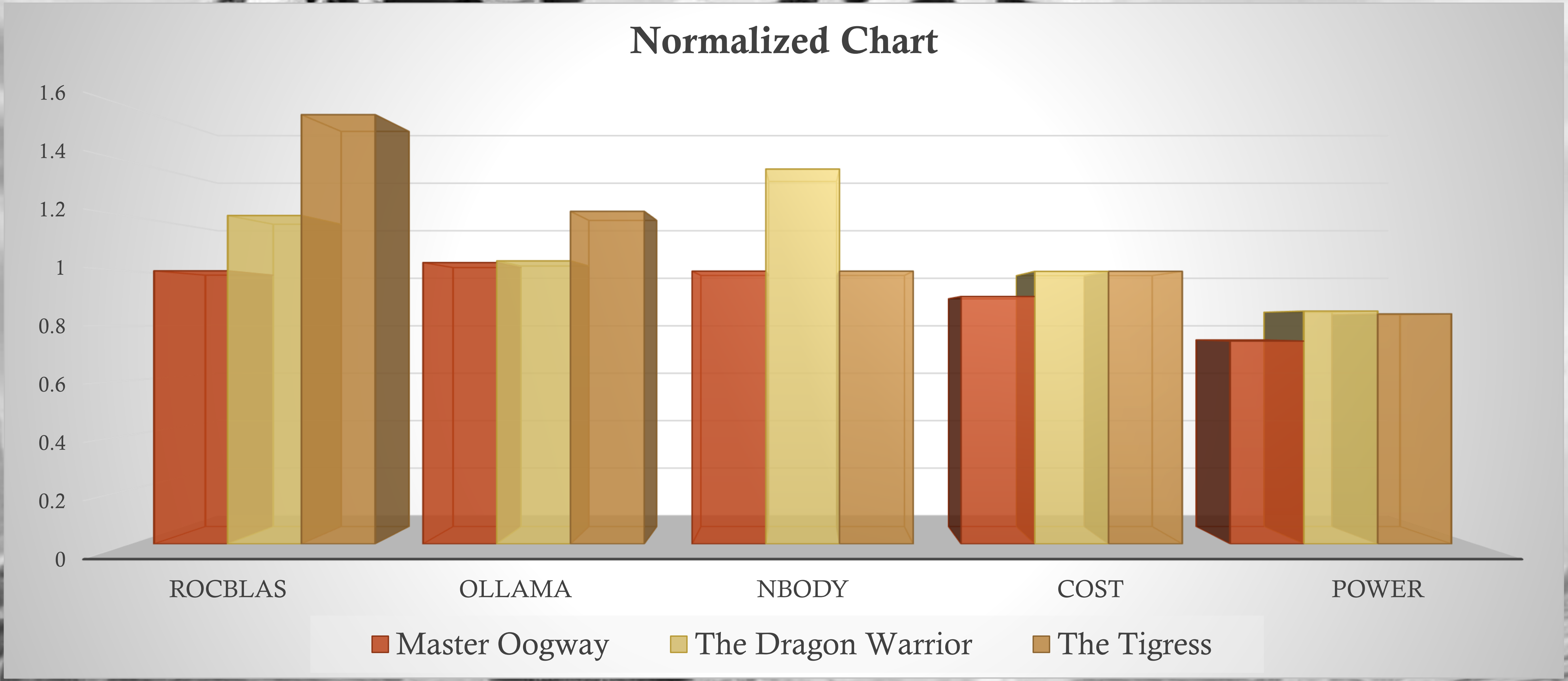
Architecture Details (KFP 6070 Gen)

RocBLAS (ExaFLOPs)	OLLaMa (Tokens/Sec)	nBody (ExaFLOPs)	FP32 MFMA (TeraFLOPs)	FP32 VALU (TeraFLOPs)	HBM Bandwidth (TB/s)
14.15	3.90 Million	1.50	4706.86	559.98	29.06

MFMA	VALU	HBM
73x	35x	3x

# of Units	Unit POWER (W)	Total Power (W)
1755	24,528	21,090,000

Architecture Overview



Budget and Pricing

# of Units	Unit Cost (USD)	Total Cost (USD)
1181	423,000	499,560,000

# of Units	Unit Cost (USD)	Total Cost (USD)
1337	411,000	549,510,000

# of Units	Unit Cost (USD)	Total Cost (USD)
1755	313,200	549,670,000

External Forces

1.4 nm process form Intel coming out within the next year or 2, along with TSMC and Samsung foundry already running 2nm process.

HBM Shortage likely to continue at least until 2027 due to its demand driven by AI infrastructure expansion.

Tariffs could raise the general prices of raw materials

Dennard Scaling is dead. Moore’s law is still “alive,” but not guaranteed to stay alive. Thus, the node shrink assumption might not be satisfied by 2030.

Next Steps

01

Facility power planning

02

Cooling and thermal management strategy

03

Network architecture design

04

Total system cost assessment

Donate to the SeaOtter Foundation

https://seaotterfoundationtrust.org/

References

- [1] “Library,” Ollama, https://ollama.com/library (accessed Feb. 15, 2026).
- [2] “Running jobs,” Running Jobs - HPC Fund documentation, https://amdresearch.github.io/hpcfund/jobs.html#large-language-models-ollama (accessed Feb. 15, 2026).
- [3] ROCm, “Releases · ROCM/Rocprofiler-Compute,” GitHub, https://github.com/ROCm/rocprofiler-compute/releases (accessed Feb. 15, 2026).
- [4] “ROCM compute profiler documentation,” ROCm Compute Profiler documentation - ROCm Compute Profiler 3.4.0 documentation, https://rocm.docs.amd.com/projects/rocprofiler-compute/en/latest/index.html (accessed Feb. 15, 2026).
- [5] “Frontier,” Oak Ridge Leadership Computing Facility, https://www.olcf.ornl.gov/olcf-resources/compute-systems/frontier/ (accessed Feb. 15, 2026).
- [6] “MI200 performance counters and metrics,” MI200 performance counters and metrics - ROCm Documentation, https://rocm.docs.amd.com/en/docs-6.0.0/conceptual/gpu-arch/mi200-performance-counters.html (accessed Feb. 15, 2026).