HPC System Proposal

Kung FLOPanda 6070 (KFP 6070)

Chief Engineers: Bryce Chen and Henry Jung

Executive Summary:

The Kung FLOPanda LLC is a limited liability company that concentrates on designing next-generation accelerator architectures for high-performance computing (HPC). The company profiles and simulates viable solutions near future through detailed analysis of available current state-of-the-art, and engineering various solutions based on requirements given.

In response to the desired requirements, there are 3 different system architectures the Kung FLOPanda LLC propose:

I. **Master Oogway:** Lowest Power and Lowest Cost, efficiency emphasis by satisfying the minimum system thresholds.
II. **Dragon Warrior:** Outputs well-rounded performances in various HPC fields with extra power consumption.
III. **The Tigress:** Maximized performance for large language models, just falling under the cost requirement.

Client's Needs or Motivation:

Scientific computing and large-scale Large Language Models require a computing platform that can sustain heavy workloads such as floating-point operations and matrix operations. Modern workloads such as particle simulations, protein unfolding, and artificial intelligence computations pushes compute units, memory, and inter node scaling to their limit, thus making a balanced architecture difficult to achieve. The Chang Corporation requires a system that delivers 9 exaflops of compute for RocBlas [3] and 1.5 exaflops for nBody benchmarks while also being able to generate 3.2 million tokens/s for llama 3 with 70 billion parameters [1]. The corporation wants to achieve these benchmarks while also staying within 25 megawatts of power and $550 million dollars. This requires difficult optimization of performance per watt and per dollar. Our company Kung FLOPanda LLC is motivated to design efficient and balanced architecture that is capable of meeting the requirements for The Chang Corporation.

Impacts:

Scientific Achievement:

A higher exascale HPC provides the computational power that is needed for the increasingly complex scientific workloads to achieve major breakthroughs and takes HPC a step closer to a zeta scale system. This processing capability allows researchers to analyze larger datasets and run increasingly complex models and simulations. An HPC system with this capability can lead to accelerating drug discoveries, cancer research, and disease treatment strategies. This type of high performance opens the door to accelerating scientific discoveries, leading researchers to address problems that were thought to be beyond the limits of modern technology.

Environmental Impact:

An increase in computational power leads to an increase in power consumption as well. We are given a 25MW constraint, but this only applies to the compute units and not cooling, networking, or facility costs. These additional energy demands have to be accounted for by Chang Corporation. Such a large power consumption leads to a significant increase in carbon emissions and environmental damage, requiring careful energy planning of the facility in order to minimize these impacts.

Societal Impact:

Advanced HPC systems can significantly impact society as it can accelerate developments in medical treatments such as drug discoveries and treatment development for diseases and cancer. This can also improve AI performance which can be used for accelerating medical imaging analysis. Not only can HPC help with the medical industry, with it being able to improve AI models researchers can use these models to optimize disaster response, energy grids, and many more leading to the improvement of society safety and efficiency.

Requirements:

The purpose of the requirements for this project is to create an HPC system that can sustain high performance across scientific computing and large-scale AI applications. The Change Corporation wants a system that can achieve at least 9.0 exaflops on rocBlas with matrix size of 16384, 3.2 million tokens per second when performing inference on Llama 3 with 70 billion parameters [2] and achieve at least 1.5 exaflops for the nBody benchmark with 1048576 particles. Power consumption is essential for HPC systems to prevent large amounts of carbon emissions and reduce the energy bill per year. The Chang Corporation wants a system to operate within 25 megawatts. This requires Kung FLOPanda LLC to make architectural decisions based off performance per watt to stay within the power constraints. A budget constraint was created for practical purposes in order to emphasize cost efficiency. The Chang Corporation wants the total cost of just the compute nodes to not exceed 550 million dollars in total. In this project the Kung FLOPanda LLC assumed two future node shrinks providing 30% reduction in power and 15% improvement of frequency per shrink, a 40% performance gain from software optimizations, and strong scaling of 100% for rocBlas, 88% for Llama 3, and 90% for nBody. The company also made some assumptions for the hardware scaling as doubling memory bandwidth increases power consumption by 20% and cost by 15%, doubling vector units increases power consumption by 40% and cost by 20%, and doubling the matrix multiply units increases power consumption by 40% and cost by 25%.

Methodology:

The current state of the art was analyzed based off profiled compute power [4] of the architecture of the accelerator inside second fastest supercomputer Frontier [5] AMD Mi250. Proper scaling was performed using Amdahl's law scaling for enhancing performances in individual processors, i.e, MFMA, VALU, etc, and also scaling for integrating more processing units into the overall system.

1. **The performance and power/cost trade offs**
   a) **HBM Memory Bandwidth**: Doubling the memory bandwidth comes with a 20% increase in power consumption and 15% increase in cost.
   b) **Vector Units (VALU):** Doubling the vector performance comes with a 40% increase in power consumption and 20% increase in cost.
   c) **Matrix Multiply Units (MFMA)**: Doubling the matrix multiply performance comes with a 40% increase in power consumption and 25% increase in cost.

Example calculation for new power and cost for 4x MFMA improvement (single unit)

New Power = $Power_{original} + \left[0.4 * Power_{original}\right] * [4-1]$

New Cost = $Cost_{original}\left[0.25 * Cost_{original}\right] * [4-1]$

2. **Calculation of new Performance for Single Unit**

$$\begin{aligned} Performance_{new} \\ &= MFMA\ improvement\ multiplier * \left[\%MFMA * Performance_{original}\right] \\ &+ VALU\ improvement\ multiplier * \left[\%VALU * Performance_{original}\right] \\ &+ HBM\ improvement\ multiplier * \left[\%HBM * Performance_{original}\right] \end{aligned}$$

First in foremost, the performance for the Mi250 on each benchmark was profiled to determine the compute to memory ratio [6]

a) **RocBLAS**: The profiled compute to memory ratio of Mi250 in RocBLAS was 86.75% MFMA, 0.223 % VALU, and 13% HBM.
b) **nBody:** The profiled compute to memory ratio of Mi250 in nBody was 0% MFMA, 99.99 % VALU, and 0.01% HBM.
c) **LLaMa**: The profiled compute to memory ratio of Mi250 in LLaMa was 65% MFMA, 0% VALU, and 35% HBM.

Example calculation for new performance of RocBLAS for 4x MFMA improvement, 2x VALU improvement, and 2x HBM improvement (single unit)

$$Performance_{new} =$$

$4 * [0.86.75 * Performance_{original}] + 2 * [0.00223 * Performance_{original}] + 2 * [0.13 * Performance_{original}]$

3. Calculation of new Performance for the system

$$Bench\ Performance_{total} = Bench\ Performance_{unit} * \#of\ units * scaling\ factor$$

Scaling Factors:

a) ***RocBLAS:*** *100%*

b) ***nBody:*** *90%*

c) ***LLaMa:*** *88%*

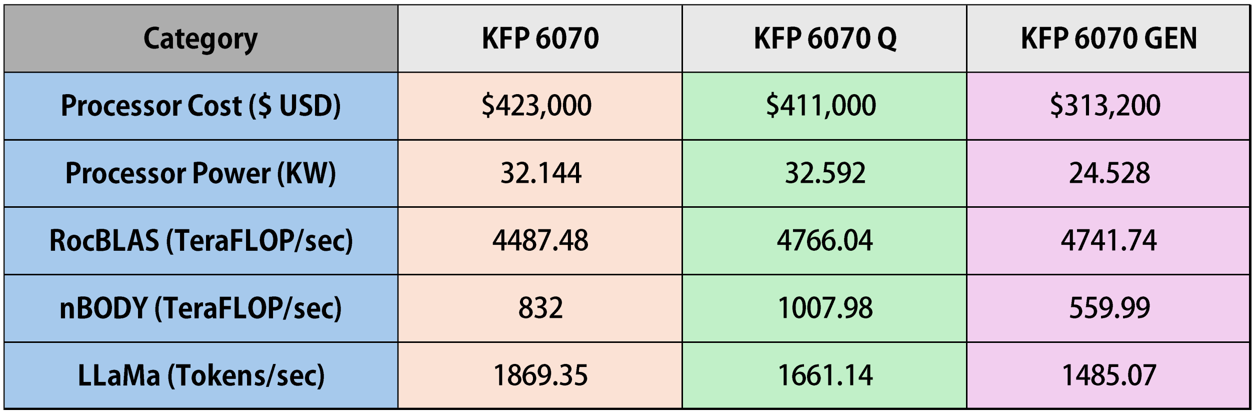
The new LLaMA performance of the system with 1000 units would be:

$$LLaMa\ Performance_{total} = LLaMa\ performance_{unit} * \#\ of\ Units * 0.88$$

Proposed HPC Solution:

This section proposed the three possible architectures, each composing the unique HPC solutions for the desired requirements for the Chang Corporation.

Table 1: Summary of Architectures



Category	KFP 6070	KFP 6070 Q	KFP 6070 GEN
Processor Cost ($ USD)	$423,000	$411,000	$313,200
Processor Power (KW)	32.144	32.592	24.528
RocBLAS (TeraFLOP/sec)	4487.48	4766.04	4741.74
nBODY (TeraFLOP/sec)	832	1007.98	559.99
LLaMa (Tokens/sec)	1869.35	1661.14	1485.07

A. Master Oogway: Efficient Cost and Power (KFP 6070)

Master Oogway is composed of a total of 1181 KFP 6070 accelerator units, and emphasizes cost and power efficiency, while meeting all the minimum requirements. The lower hardware cost allows extra budgeting room for other fixed costs such as cooling and networking. This system is optimized for clients seeking cost-effective options.

Table 2: Summary of Master Oogway HPC System

Category	Actual Threshold	Target Threshold
System Cost ($ USD)	$499.6 Million	$550 Million
Power Consumption (MW)	18.6	25
RocBLAS (ExaFLOP/sec)	9.01	9
nBODY (ExaFLOP/sec)	1.5	1.5
LLaMa (Tokens/sec)	3.3 Million	3.2 Million

B. Dragon Warrior: Well-Rounded (KFP 6070 Q)

Dragon Warrior is composed of a total of 1337 KFP 6070 Q accelerator units, outputting optimized performances for all 3 Benchmarks. The accelerator unit is the most expensive and the power consumption is the biggest out of the 3 possible options, with the cost coming in for a trade with performance enhancement in all 3 corners. This system becomes the best option for clients seeking an HPC system that practice balance but also peak performance for more cost.

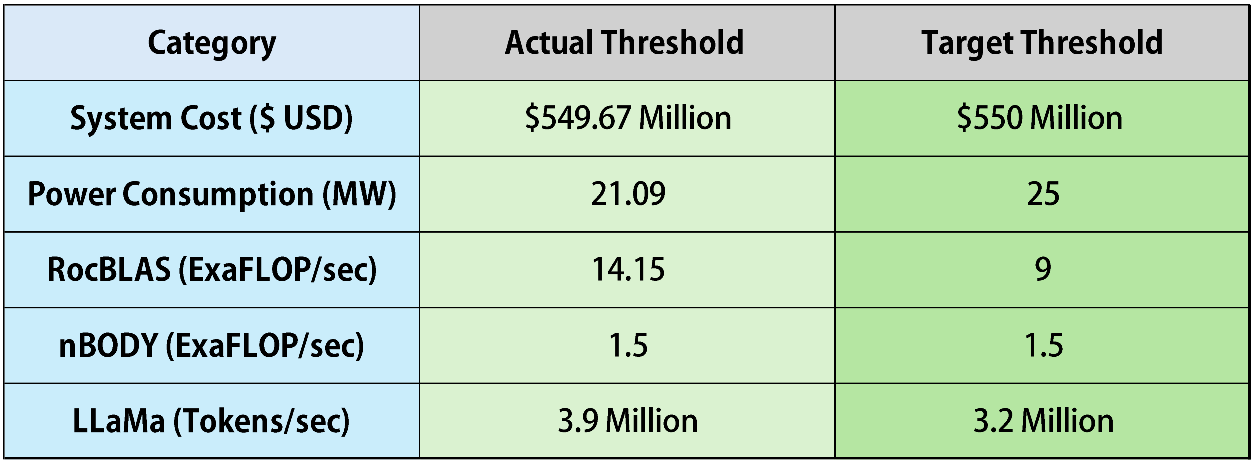
Table 3: Summary of Dragon Warrior HPC System

Category	Actual Threshold	Target Threshold
System Cost ($ USD)	$549.51 Million	$550 Million
Power Consumption (MW)	21.35	25
RocBLAS (ExaFLOP/sec)	10.83	9
nBODY (ExaFLOP/sec)	2.06	1.5
LLaMa (Tokens/sec)	3.32 Million	3.2 Million

C. The Tigress: LLM Optimized. (KFP 6070 Gen)

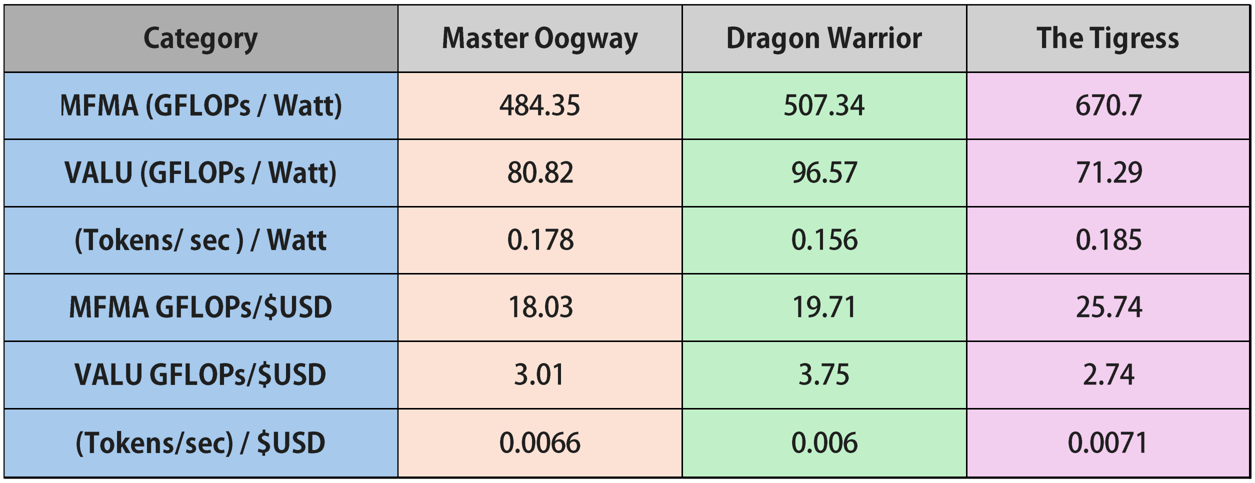
The Tigress, composed with total of 1755 KFP 6070 Gen units, has a strong design and development focus for Large Language Models (LLM), with maximized performances in MFMA operations with enough memory bandwidth improvement. There is a strong focus on improvement in RocBLAS up to 14.15 ExaFLOP/sec, and maximized LLaMa with 3.9 million token generations/sec. The increases system power and cost comes with shifting all its capabilities to suit LLM.

Table 4: Summary of The Tigress HPC System



Category	Actual Threshold	Target Threshold
System Cost ($ USD)	$549.67 Million	$550 Million
Power Consumption (MW)	21.09	25
RocBLAS (ExaFLOP/sec)	14.15	9
nBODY (ExaFLOP/sec)	1.5	1.5
LLaMa (Tokens/sec)	3.9 Million	3.2 Million

Table 5: Summary of cost and power per GFLOP ($1x10^9$)/second



Category	Master Oogway	Dragon Warrior	The Tigress
MFMA (GFLOPs / Watt)	484.35	507.34	670.7
VALU (GFLOPs / Watt)	80.82	96.57	71.29
(Tokens/ sec) / Watt	0.178	0.156	0.185
MFMA GFLOPs/$USD	18.03	19.71	25.74
VALU GFLOPs/$USD	3.01	3.75	2.74
(Tokens/sec) / $USD	0.0066	0.006	0.0071

Budget and Pricing:

1. Master Oogway

The Master Oogway contains the emphasis on efficiency with minimum resources used. The total projected hardware cost alone is to be $499.6 Million which leaves noticeable headroom under the $550 Million threshold. This headroom should partially account for the other fixed costs that come along with HPC systems such as cooling, and power for networking.

2. Dragon Warrior

The Dragon Warrior follows the well-rounded principle, providing balanced but also optimized performances in various benchmarks, with the higher cost of $549.51 Million. The cost comes just below the limit, and it is driven by more number of accelerators with general performance bumps.

3. The Tigress

Likewise, The Tigress also comes with a higher cost of $549.67 Million. This increased cost is driven by strong focus on improvement in Matrix Fused Multiply Add (MFMA) operations with a higher performance to cost trade-off margin.

External Forces:

HPC systems require high memory bandwidth and high-bandwidth memory (HBM) are key components for reducing the negatives of the memory wall. With sky rocketing demand driven for more HBM due to construction of data centers and increasing number of AI infrastructures in the world, the HBM supply stability until the year 2030 remains uncertain. With only few companies like SK Hynix, Samsung Electronics, and Micron Technology available for high-quality production, the cost increase in HBM is possible, which also puts the achievable cost requirement of $550 million to a question.

With slowdown of Moore's law and the end of Dennard Scaling, the assumption of improvement if power efficiency and switching speed of transistors are not guaranteed to match the actual advancement in the year 2030, especially with current technology already providing single nanometer scale fabrication process.

Conclusions:

The presented architectures demonstrates that it is possible to achieve all the performance, power and cost requirements that were set by the Chang Corporation and provides them with multiple pathways for their next generation HPC system. There are still some risks and open challenges that remain before building this new HPC system. The 25-megawatt constraint only includes the consumption of just the compute nodes excluding the cooling system, networking infrastructure, and the facility operations. This introduces additional power, thermal, and cost considerations that has to be addressed before moving on to the next stages. The proposed architectures rely on assumptions of process node shrinks, software optimization projections, and close to ideal scaling behavior which can potentially change under real world operating conditions.

As a next step, a power modeling system needs to be created that includes cooling, networking, and facility integration in order to have more of an accurate total energy consumption of the system. Engaging with energy providers and infrastructure companies will be crucial to develop an efficient and environmentally reasonable power strategy that is capable of powering this entire system. By addressing these risks, the Chang Corporation can successfully choose architecture, successfully build the system, and establish themselves as the leading Supercomputer in the world.

References & Appendix:

[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).

[7]

- **Minimum** of 9 Exascale (9x10^{18}) Floating Point Operations Per Second (FLOPs) for LINPACK Benchmark
- **Minimum** of 3.2 million Token Generations Per Second for Large Language Model based on LLaMa Benchmark
- **Minimum** of 1.5 Exascale (1.5x10^{18}) FLOPs for particle collision simulation using nBody Benchmark.
- **Maximum** power consumption of 25,000,000 Watts, (25MW).
- **Maximum** hardware development cost of $ 550,000,000 USD, ($ 550 Million).

Statement of Work:

This is the delegation of work listed below:

• Benchmark Profiling:	Bryce-90; Henry-10;
• Spreadsheet Setup:	Bryce-20; Henry-80;
• Solution Finding & Justification:	Bryce-40; Henry-60;
• Paper Writeup:	Bryce-50; Henry-50;