Tolaga Research
Harness the Power of Intelligence

Research Note
November 2026

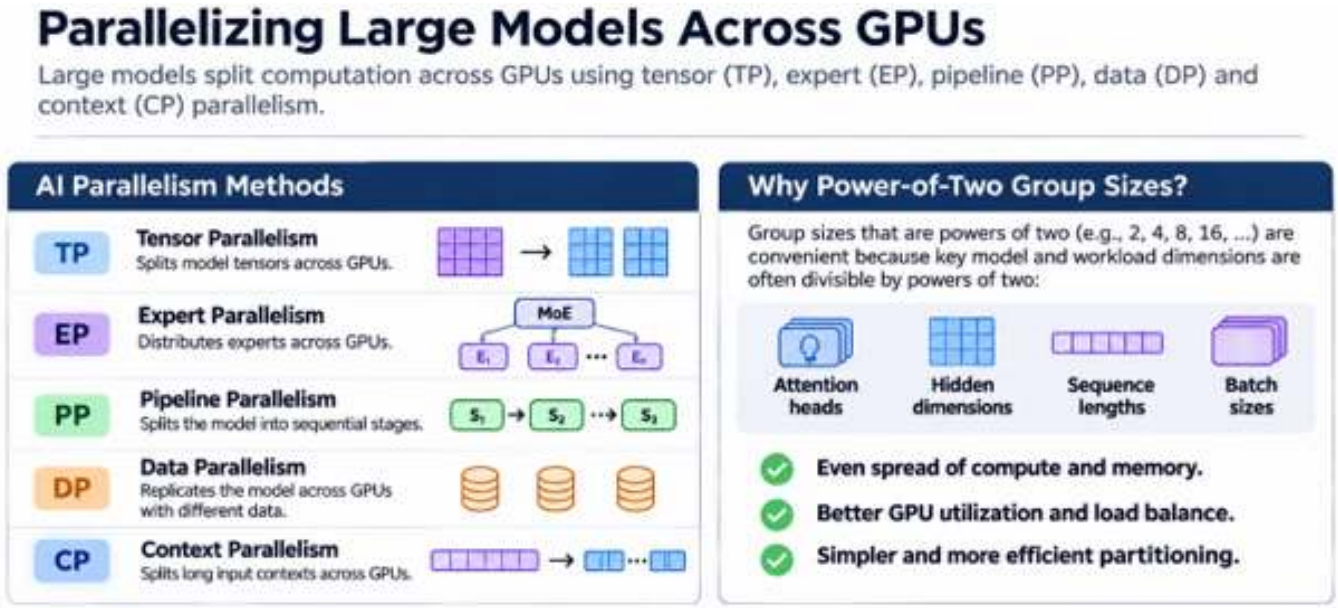
Why Powers of Two Show Up in AI GPU Clusters

Phil Marshall PhD, Chief Research Officer, Tolaga Research



Model parallelism and workload partitioning

Large models distribute computation across GPUs using tensor (TP), expert (EP), pipeline (PP), data (DP) and context (CP) parallelism. Power-of-two group sizes are convenient because model dimensions, attention heads, expert counts, sequence lengths and batch sizes are often divisible by powers of two, enabling balanced computation and memory allocation.



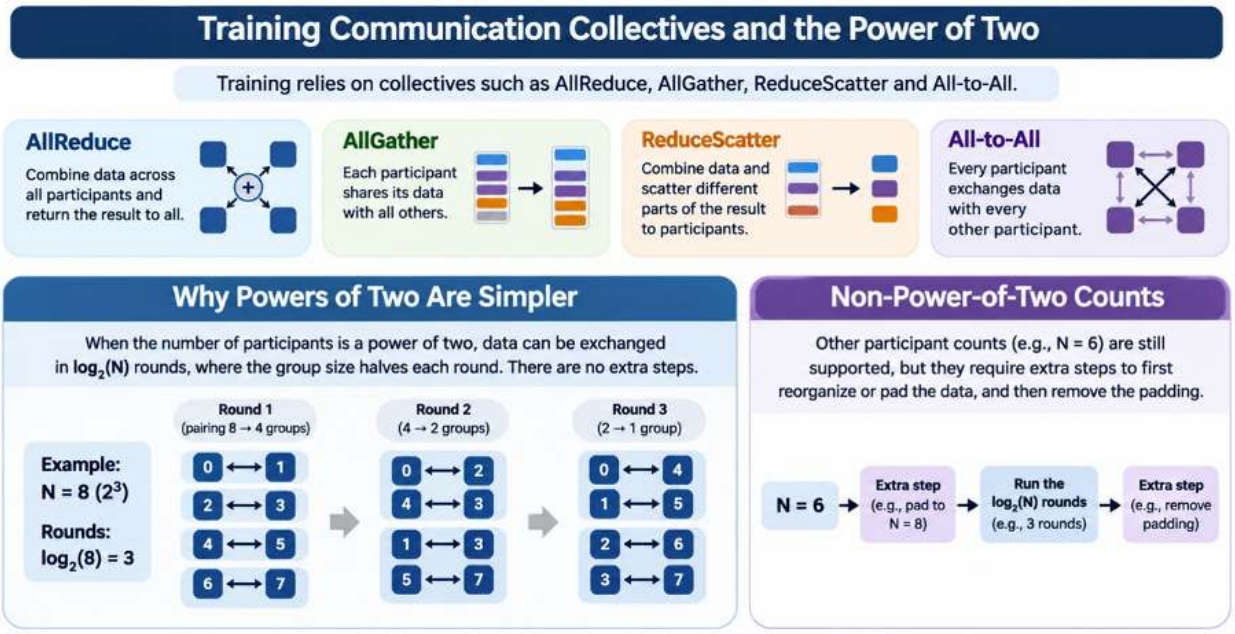
For example, TP=8, PP=4 and DP=16 requires 8 × 4 × 16 = 512 GPUs. Each factor divides cleanly into the model's dimensions and into the hardware, which keeps scheduling simple and workloads balanced.

However, power-of-two configurations are not mandatory. Llama 3.1 405B has 126 transformer layers, which divide evenly across three or six pipeline stages, but not four or eight. Similarly, GPT-3's 96 attention heads and Kimi K2's 384 experts are not powers of two, yet both are divisible by several power-of-two factors. These examples illustrate that power-of-two parallelism remains convenient even when model dimensions are not powers of two, but efficient partitioning ultimately depends on model divisibility and workload balance.

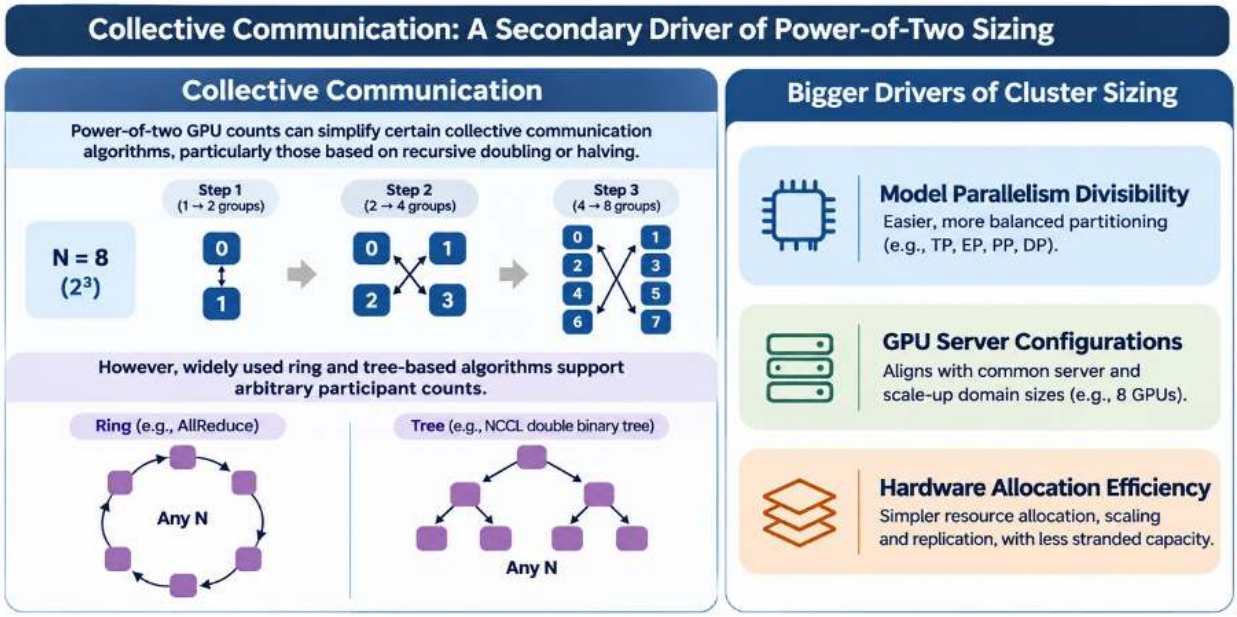
© 2026 Tolaga Research. All rights reserved.

Collective communications

Training relies on collective operations such as AllReduce, AllGather, ReduceScatter and All-to-All to exchange gradients, model parameters and intermediate results between GPUs. These operations synchronize model updates and coordinate distributed computation.



Certain algorithms, particularly recursive doubling and halving, progressively exchange and aggregate data across GPU groups, requiring a logarithmic number of communication stages when participant counts are powers of two. Non-power-of-two participant counts may require additional processing, although widely used ring and tree-based algorithms support arbitrary counts.



© 2026 Tolaga Research. All rights reserved.

Collective communication is therefore generally a secondary consideration in power-of-two GPU allocation. Actual performance depends on message size, network bandwidth, topology, congestion and the chosen communication algorithm.

Physical versus logical capacity: NVL72 example

NVIDIA's NVL72 architecture illustrates how physical GPU configurations can differ from logical training allocations. Although each NVLink domain contains 72 GPUs, power-of-two allocations such as 64 GPUs are common in AI training. Meta's Prometheus supercluster provides a concrete example, allocating 64 GPUs per domain to training and retaining the remaining eight as spares. Meta attributes this configuration to the efficiency of power-of-two parallelism groups, with the spare capacity providing an additional fault-tolerance benefit.

Inference Parallelism and Resource Allocation

Tensor, expert and context parallelism in inference can also benefit from power-of-two group sizes. However, inference generally offers greater allocation flexibility than training because deployments can run multiple independent model replicas, each configured to use available hardware efficiently.

Long-context inference introduces additional constraints. Larger key-value (KV) caches increase memory requirements, while context parallelism distributes long sequences across GPUs and introduces additional communication. These factors influence the number and placement of GPUs required per replica, potentially favoring larger parallelism groups. Nevertheless, the ability to deploy independent replicas provides greater flexibility in allocating physical GPU capacity without requiring uniform power-of-two configurations.

Conclusions

- **Logical parallelism:** Powers of two simplify model partitioning and load balancing, but are not mandatory.
- **Collective communication:** Some algorithms favor power-of-two participant counts, while ring and tree-based alternatives support arbitrary sizes.
- **Physical infrastructure:** Servers, switching fabrics and NVLink domains do not inherently require power-of-two GPU counts, as NVL72 demonstrates.
- **Resource allocation:** Power-of-two workload groups can coexist within larger clusters of arbitrary size, with inference replicas providing additional allocation flexibility.

Power-of-two configurations remain common because they align conveniently with model dimensions and parallelism strategies. However, the practical objective is to minimize workload imbalance, communication overhead and stranded GPU capacity, rather than adhere to power-of-two conventions. The optimal allocation depends on model architecture, parallelism requirements, network topology and fault-tolerance objectives.

© 2026 Tolaga Research. All rights reserved.

About the Author: Phil Marshall, PhD

Phil Marshall's work in AI began more than 30 years ago with Kohonen maps, discrete neural networks, machine learning, and fuzzy logic. More recently, his work has targeted AIenabled solutions for the automotive, healthcare, manufacturing, and smart city sectors. Over the past decade, he has applied natural language processing and large language models to analyze online content and generate distinctive, datadriven industry research insights.

Marshall also brings deep experience in telecom, wireless, mobile, and networking technologies, fields he has worked in since the early 1990s. More recently, he has combined this infrastructure background with advanced software and modeling capabilities to develop Tolaga's AI Infrastructure Simulator, which is used to analyze the performance, cost, and scalability of emerging AI architectures.

Marshall is a dual citizen of the United States and New Zealand and currently resides in Auckland, New Zealand.

Contact: philip.marshall@tolaga.com

© 2026 Tolaga Research. All rights reserved.