

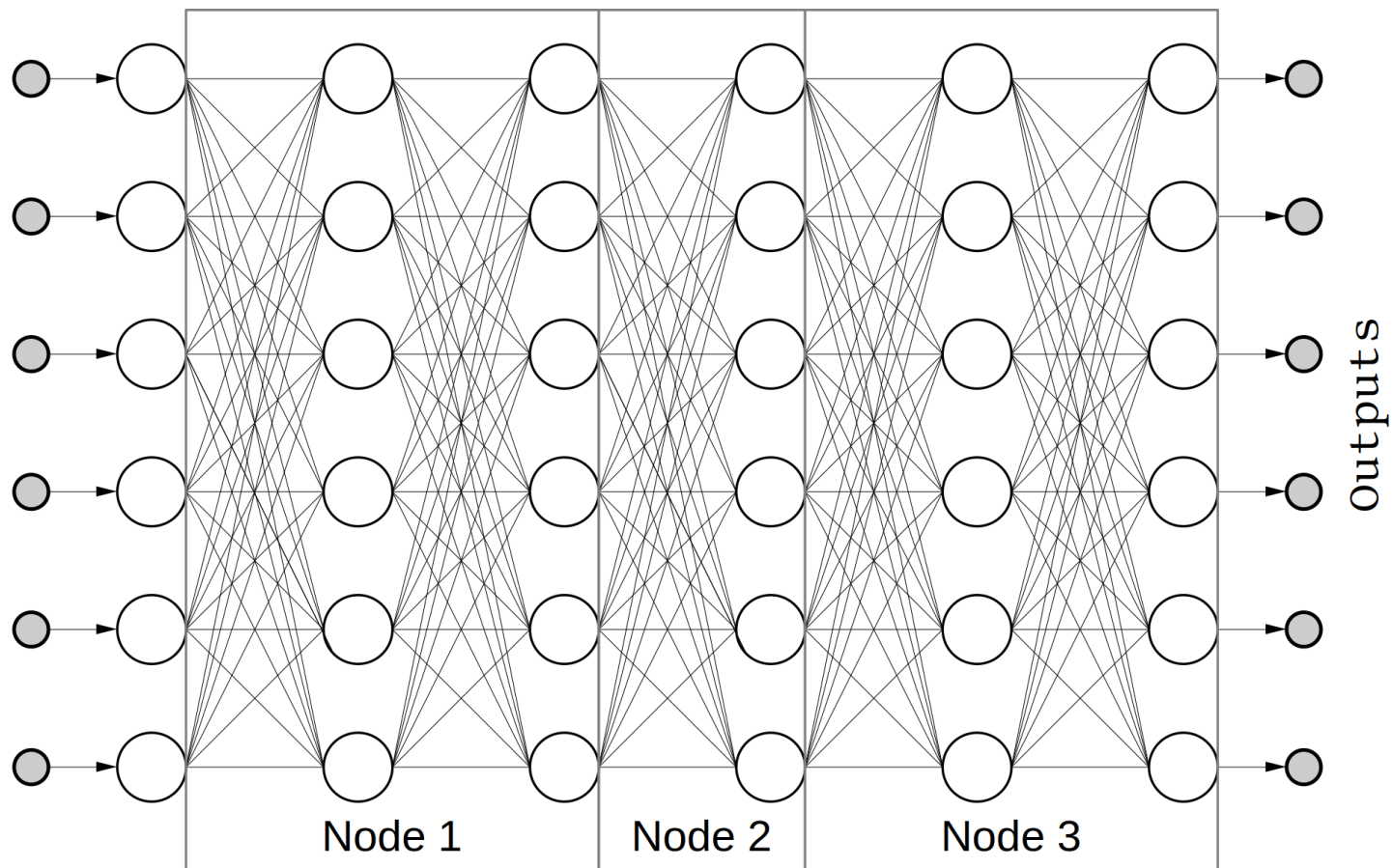
Practical Integration of Heterogeneous Resources for LLM Inference

Nikolay Khrapov, IITP RAS, Moscow, Russia

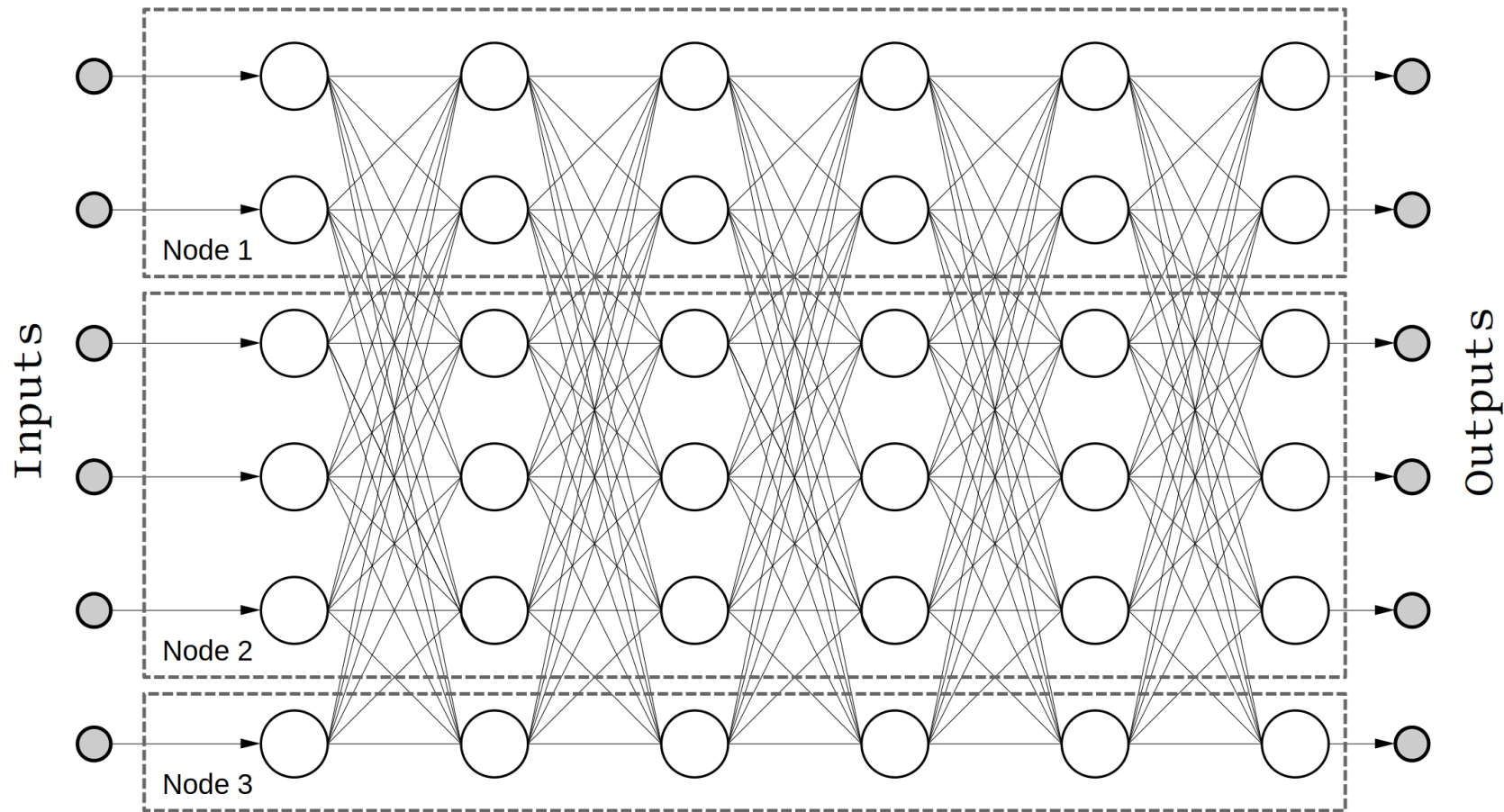
Практическая интеграция разнородных ресурсов для запуска LLM

Храпов Николай, ИППИ РАН. Москва, Россия

Vertical decomposition



Tensor decomposition



Resource

- **GPU + VRAM.** Video memory capacity and data transfer speed are of critical importance.
- **CPU + RAM** (CPU cores, RAM volume).
- **Storage** (The frequent launching of new models requires high data transfer speeds to and from the disk. SSD is critical for research.)

Software

- **vLLM.** The vLLM project is a high-performance LLM servicing system optimized for production deployment on clusters of powerful GPUs.
- **Exo.** Exo approach: running LLM on every-day personal devices—iPhones, iPads, Android smartphones, laptops—forming a peer-to-peer cluster.
- **NVIDIA Personal AI Router (PAIR)** acts as a routing layer that automatically discovers available nodes and distributes independent AI requests across them, allowing multi-agent workflows to leverage idle compute without requiring complex setup or cloud services.

llama.cpp

- **CPU & GPU Support:** It can run purely on the CPU or leverage graphics cards via CUDA, Metal, Vulkan, or ROCm for hardware acceleration.
- **GGUF Format Support:** It uses a single, compact file format that contains both the model weights and necessary metadata (tokenizer, configuration).
- **Quantization (Compression):** It compresses model weights down to 2–8 bits, allowing large AI models to run on consumer-grade laptops and PCs with minimal quality loss.
- **Cross-Platform:** It compiles easily and runs on Windows, Linux, macOS, and ARM devices like the Raspberry Pi.
- **Built-in Server:** The llama-server utility launches a local HTTP server with an OpenAI-compatible API, making it easy to connect to web interfaces like Open WebUI.
- **Distributed execution:** it is possible to distribute the LLM across the CPUs and GPUs of different machines using sequential partitioning.

Computer 1:

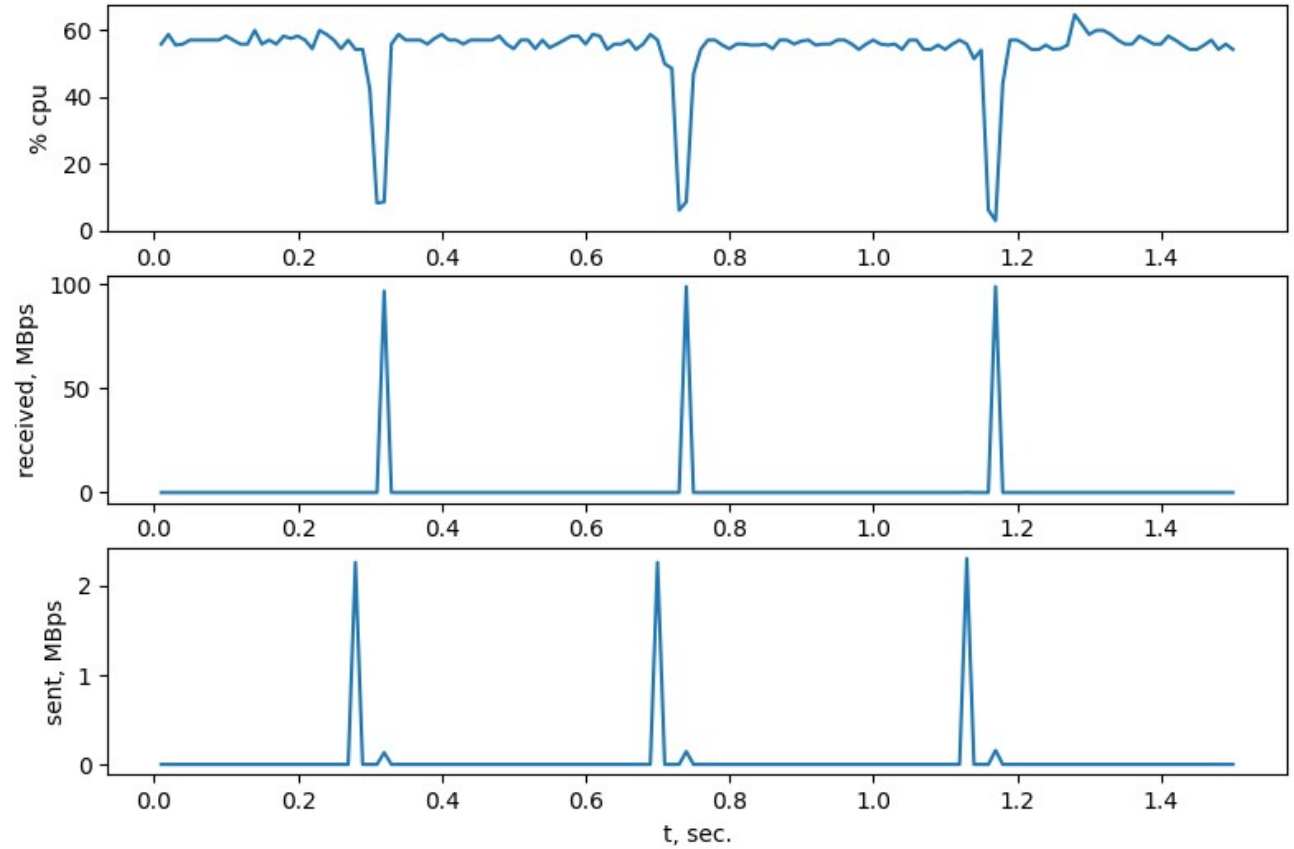
- Amd Ryzen 9 7945hx;
- 16 cores;
- 64 GB RAM (5200 Mhz).
- Nvidia 4060M, 8GB.

Computer 2:

- Intel Core i7;
- 32 GB RAM (2133 MHz);
- emdedding graphics;
- Nvidia 4Gb;
- use only GPU memeory.

Network connection: 1 Gbit/sec.

Latency: 1ms.



Computer 1:

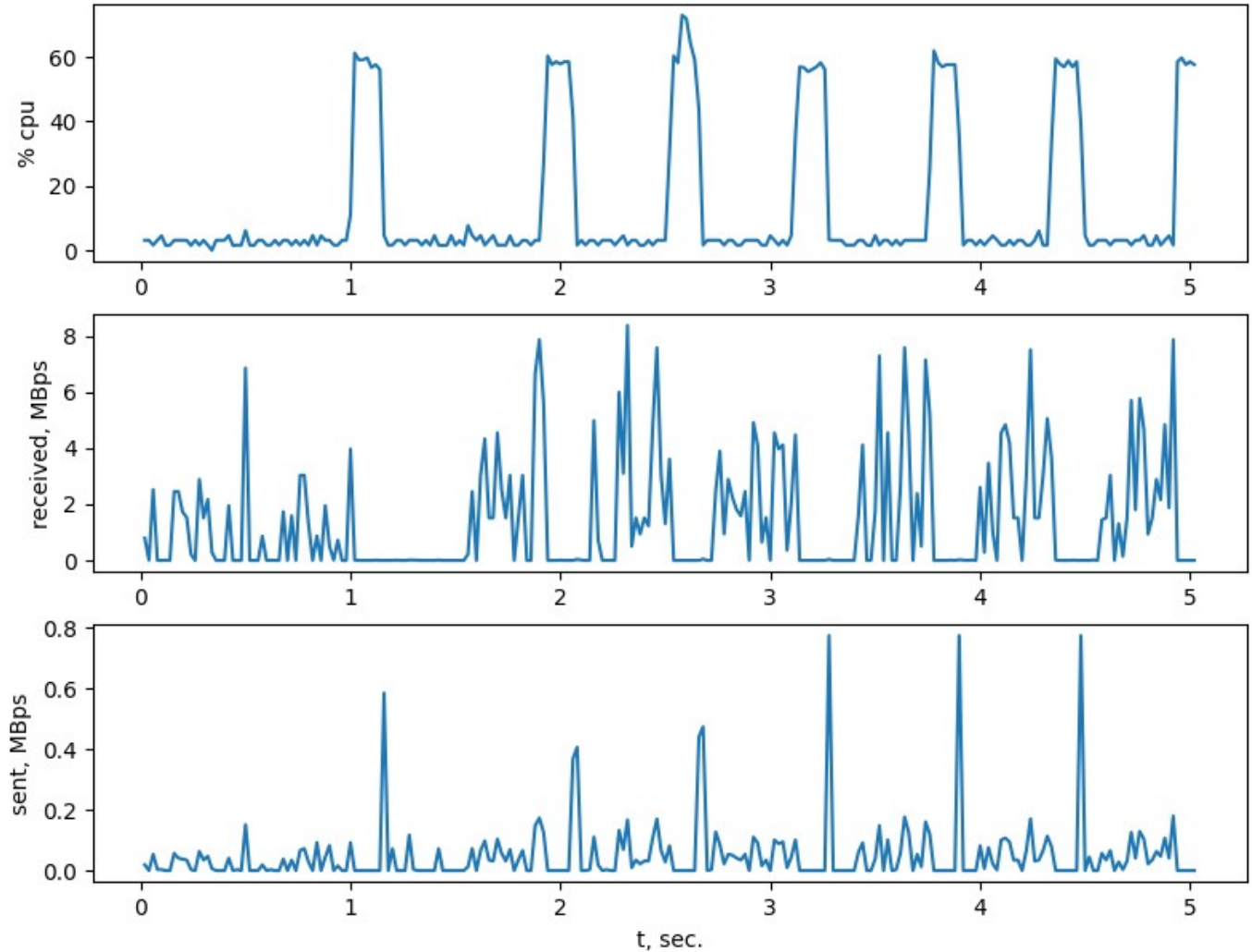
- Amd Ryzen 9 7945hx;
- 16 cores;
- 64 GB RAM (5200 Mhz).
- Nvidia 4060M, 8GB.

Computer 2:

- Intel Core i7;
- 32 GB RAM (2133 MHz);
- emdedding graphics;
- Nvidia 4Gb;
- use only GPU memeory.

Network connection: 54 Mbit/sec.

Latency: 10ms.



Computer 1:

- Amd Ryzen 9 7945hx;
- 16 cores;
- 64 GB RAM (5200 Mhz).
- Nvidia 4060M, 8GB.

Computer 2:

- Intel Core i7;
- 32 GB RAM (2133 MHz);
- emdedding graphics;
- Nvidia 4Gb;
- use only GPU memeeory.

Computer 3:

- Amd Ryzen 7 8845hx;
- 8 cores;
- 32 GB RAM (6000 Mhz).

Network connection: Gbit/sec.

Latency: 1ms.

