

Mohamed Ghamri, Marc Lacoste, Divi De Lacour

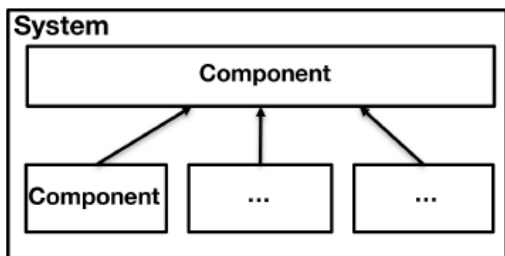
Orange Innovation, France

Challenge and Approach

- **Large-scale AI systems:** limited resources + security + performance?
- **Disaggregation:** modularization of big systems
- Explored separately for AI, HW and security

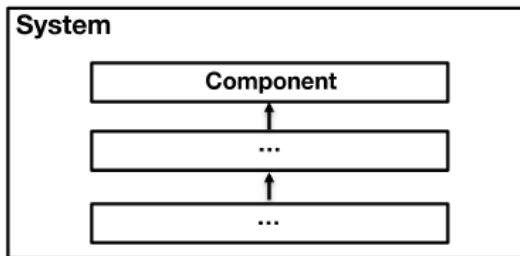
Disaggregation Patterns

Horizontal Disaggregation (HD)



- + scalability, latency, throughput
- synchronization attack surface

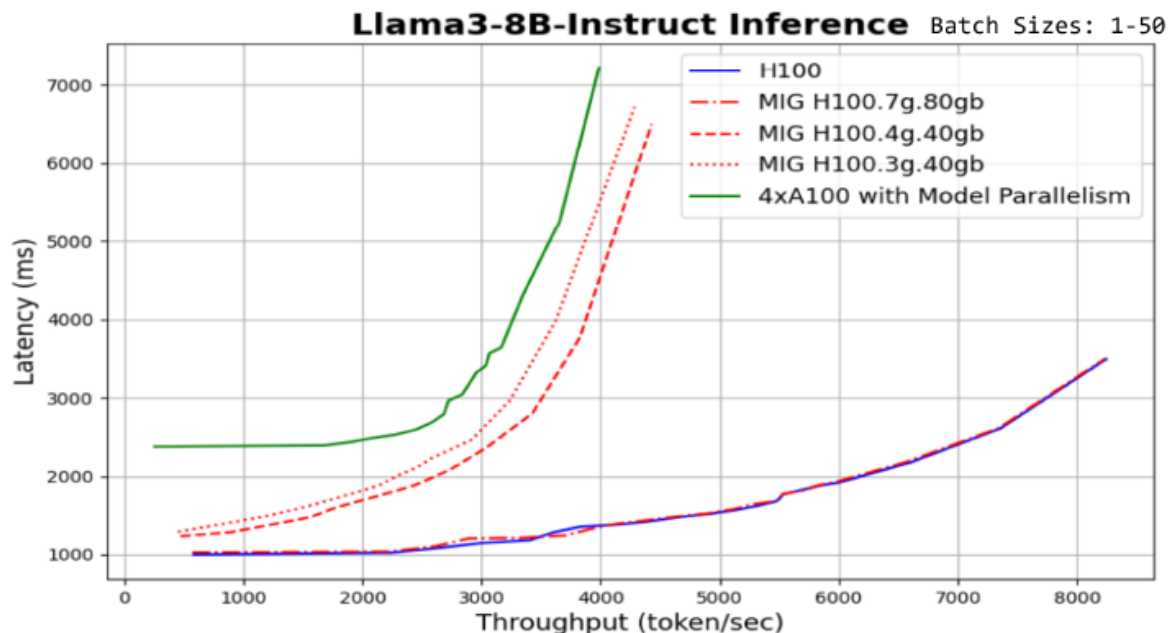
Vertical Disaggregation (VD)



- + layer optimization isolation
- performance

HD vs. VD? LLMs, federated learning, ...

Experimental Results: LLM Use-Case



Single GPU: CPU: 2xIntel Xeon Gold 6448Y 32 cores, GPU: NVIDIA HGX 4xH100 SXM 80GB, 512GB RAM
 Networked GPUs: CPU: 2xAMD EPYC 7252 16 cores, GPU: NVIDIA 8xA100 40GB, 512GB RAM

- Which HW disaggregation scales better?
 1. **single GPU** best performance
 2. **sliced GPU (MIG)** small overhead but flexibility
 3. **networked GPUs** large network overhead
- Next steps: confirm findings on FL

contact: marc.lacoste@orange.com