OPTIMUSNIC: Offloading Optimizer State to SmartNICs for Efficient Large-Scale AI Training

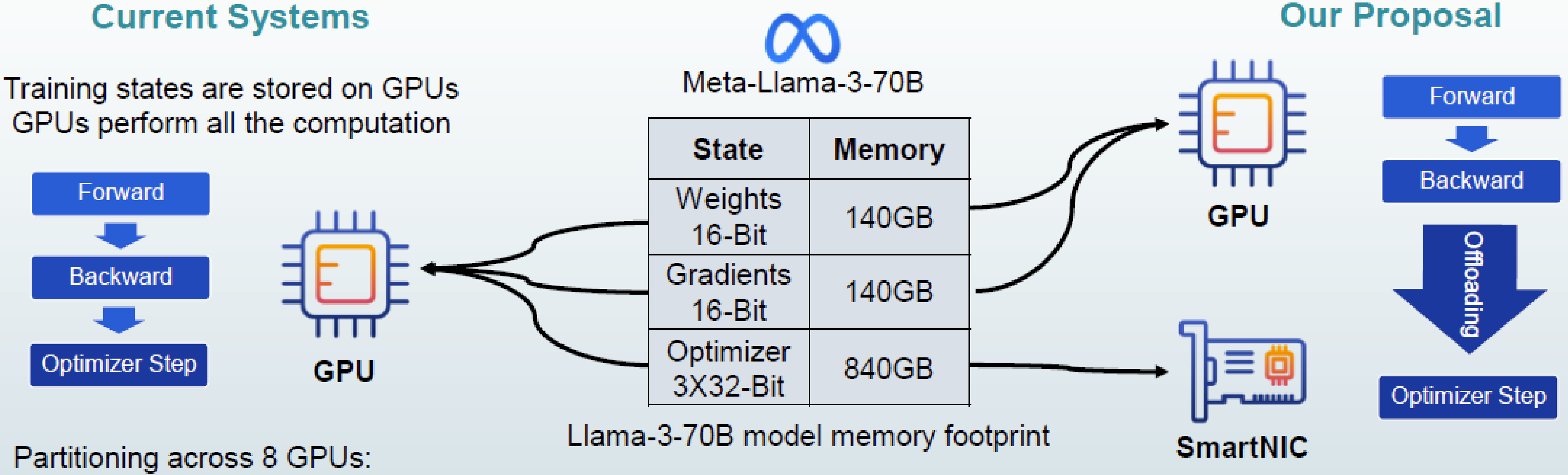
Achref Rebai Marco Canini
King Abdullah University of Science and Technology, Saudi Arabia

Problem Statement

Current Systems

Training states are stored on GPUs
GPUs perform all the computation

Our Proposal



State	Memory
Weights 16-Bit	140GB
Gradients 16-Bit	140GB
Optimizer 3X32-Bit	840GB

Llama-3-70B model memory footprint

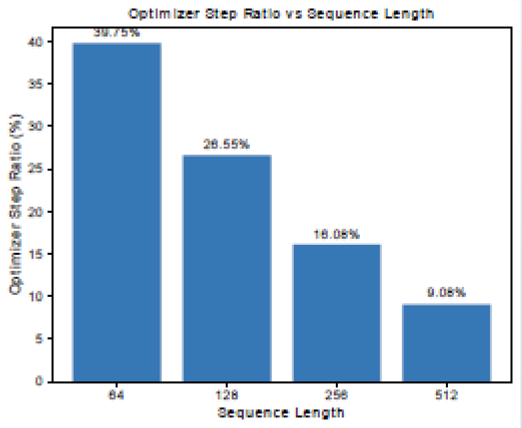
Total memory footprint: **1120 GB!**

Partitioning across 8 GPUs:
140GB per GPU >> A100 or H100!
OOM!

On GPU model memory footprint is reduced by 4X

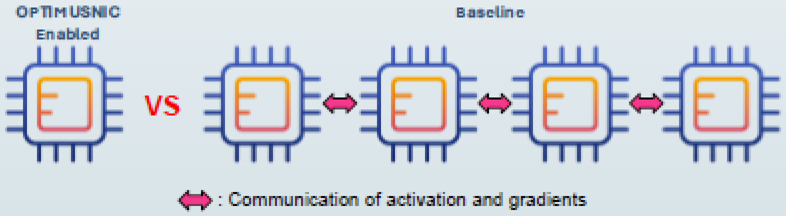
Sparing GPU Cycles

- Optimizer overhead is non-negligible.
- SmartNIC offloading frees GPU cycles and overlapping operations accelerates training iteration.



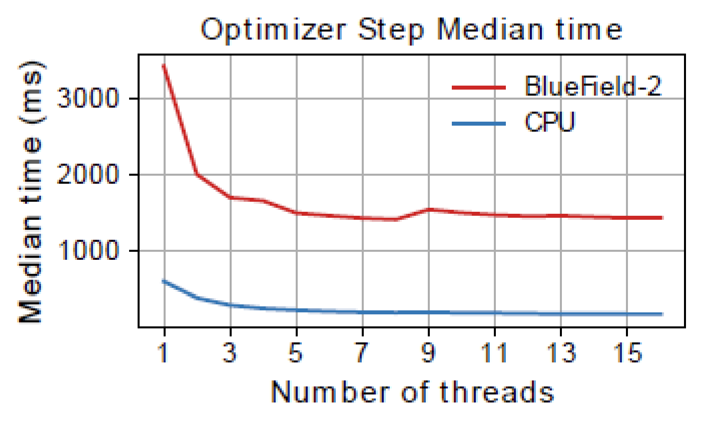
Less Communication Volume

- GPUs hold more model layers → fewer activation and gradient transfers.
- Model parallelism across GPUs slows down forward and backward passes by up to 7.8% and 5%, respectively.



Nvidia BlueField SmartNICs

- Distributing the load across multiple SmartNICs can achieve parity.
- BlueField-3 is a great candidate for this approach!


Optimizer Step Median time
Median time (ms)
BlueField-2
CPU
Number of threads