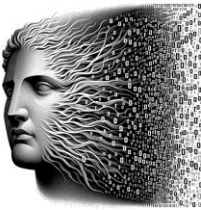
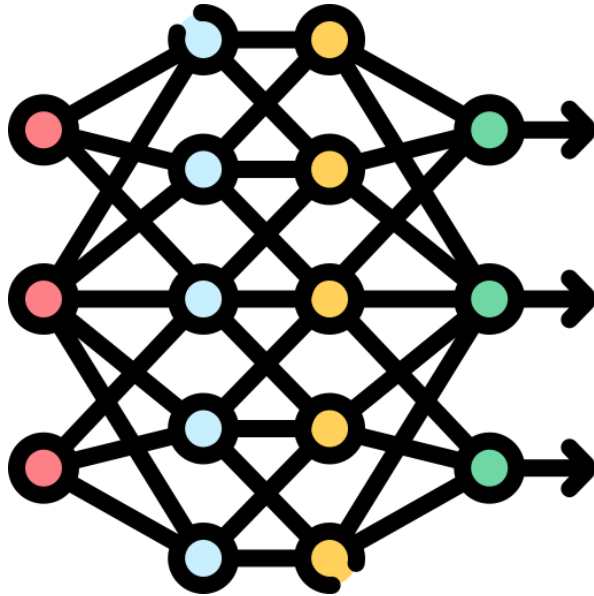


Atalanta: A Bit is Worth a “Thousand” Tensor Values

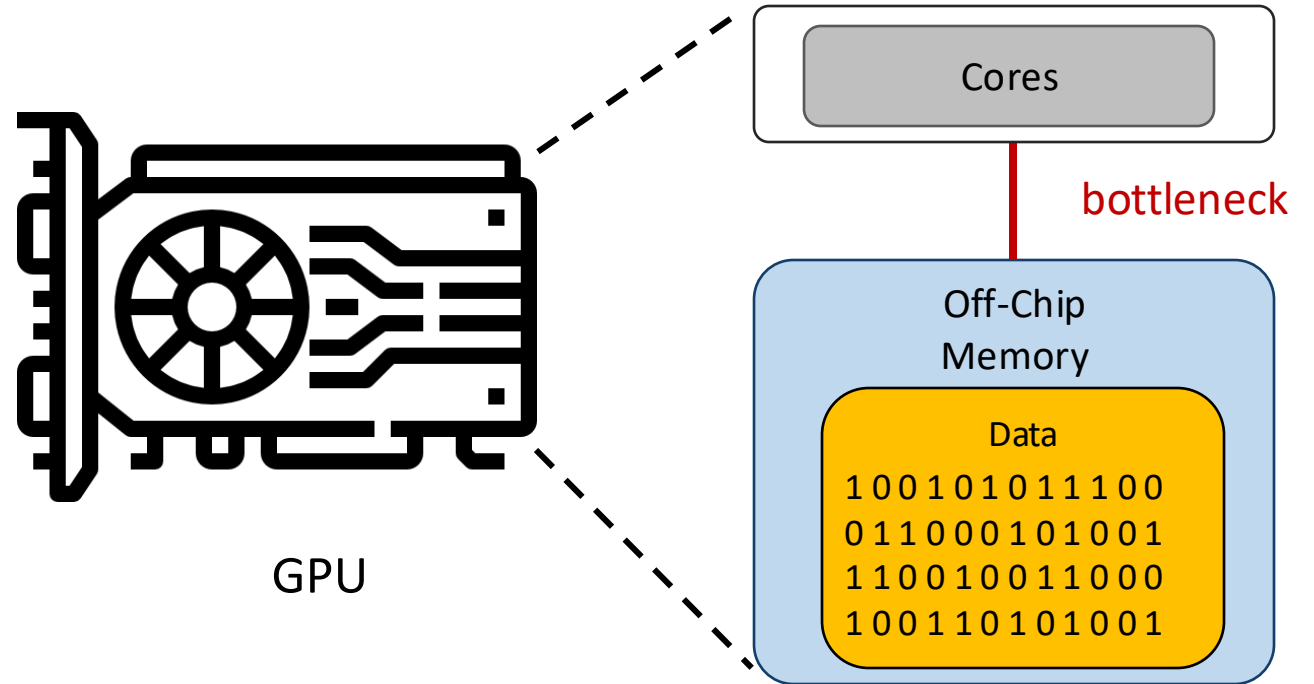
Alberto Delmas Lascorz, Mostafa Mahmoud, Ali Hadi Zadeh,
Miloš Nikolić, Kareem Ibrahim, [Christina Giannoula](#),
Ameer Abdelhadi, Andreas Moshovos



Growing Memory Footprint of Deep Learning Models

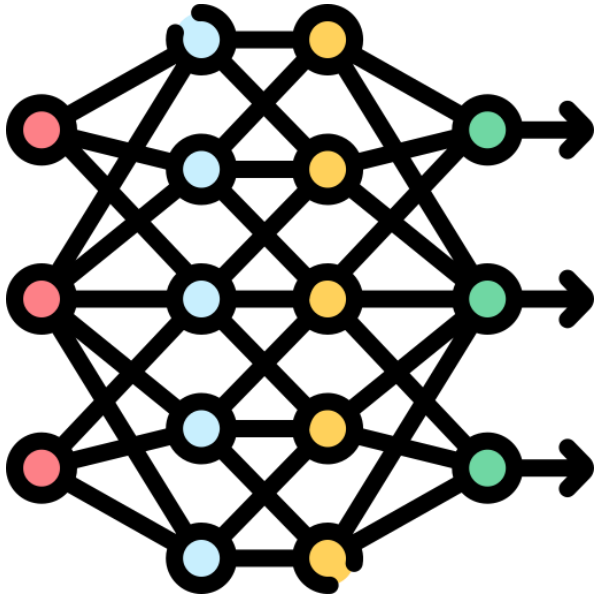


E.g., Llama2: 7B Params, 26GB Mem

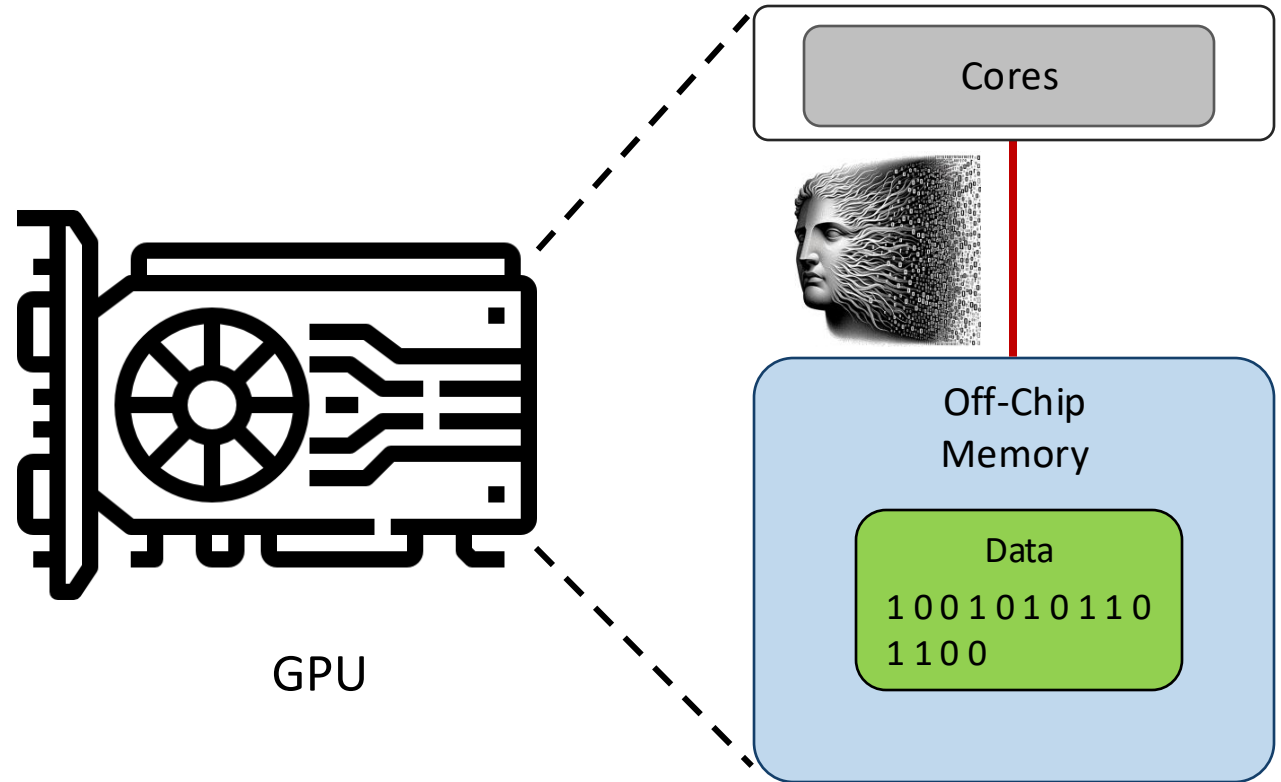


Key Limitation:
Expensive Tensor Data Accesses from Off-Chip Memory

Atalanta: Overview



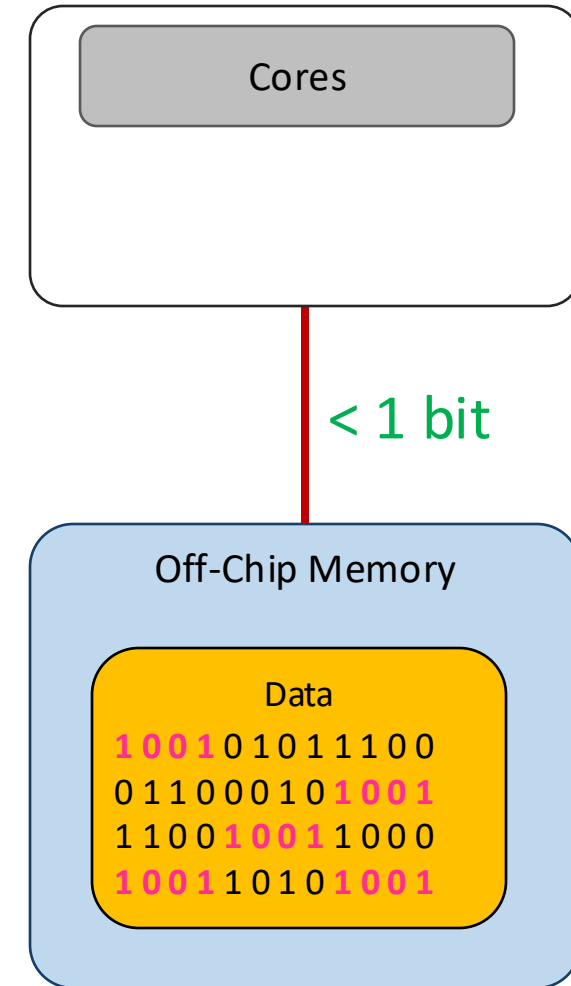
E.g., Llama2: 7B Params, 26GB Mem



Atalanta:
A Near-Optimal Compression Engine for Deep Learning Models

Atalanta: Brief Overview

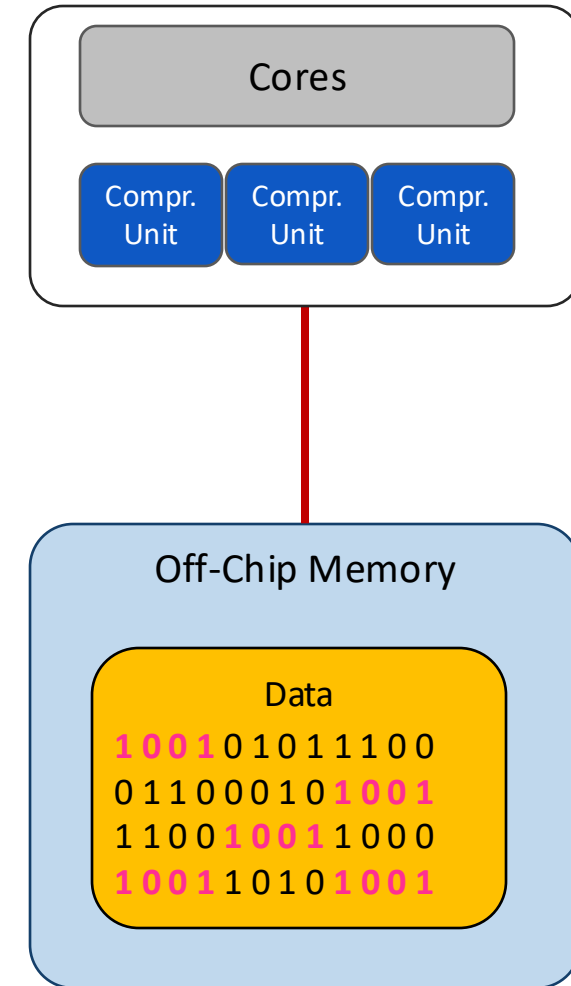
- Frequent values need **less than 1 bit**
 - Arithmetic Coding [Langdon, 1984]



Atalanta: Brief Overview

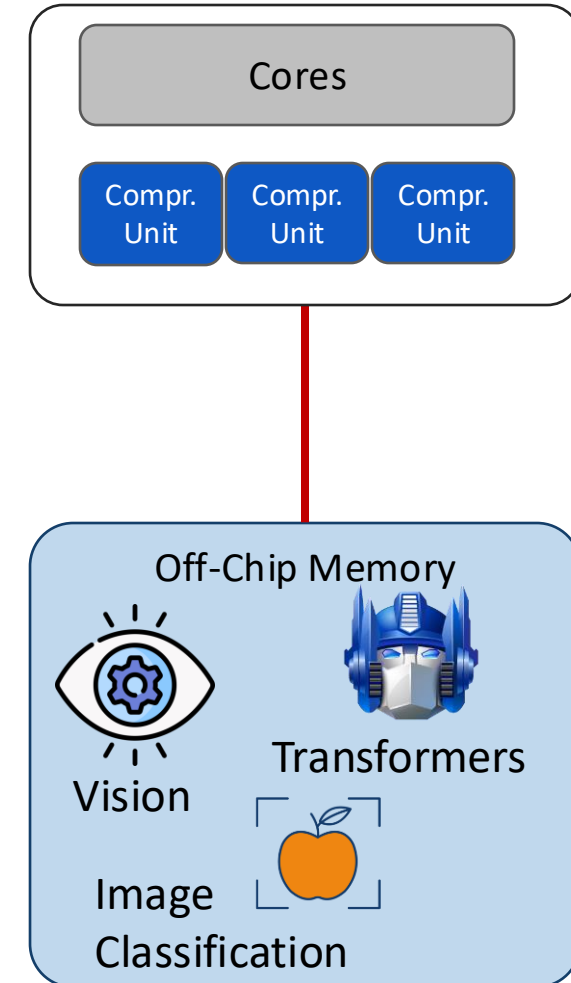
- Frequent values need less than 1 bit
 - Arithmetic Coding [Langdon, 1984]
- Low Hardware Costs

< 50
bytes



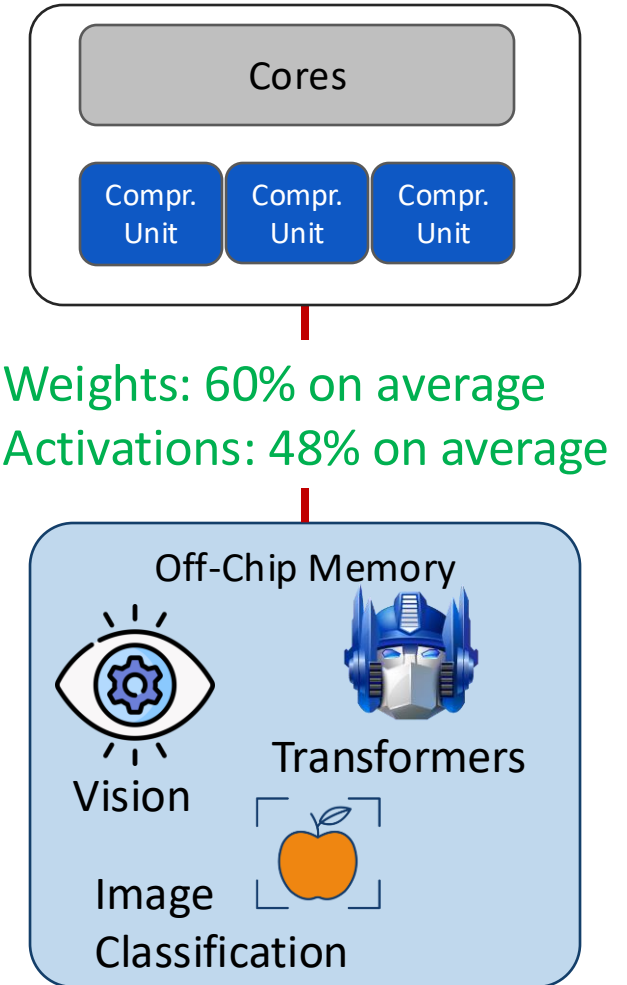
Atalanta: Brief Overview

- Frequent values need less than 1 bit
 - Arithmetic Coding [Langdon, 1984]
- Low Hardware Costs
- Programmer **Transparent**
- **Lossless**
- **Inference** + **Training** (forward pass)



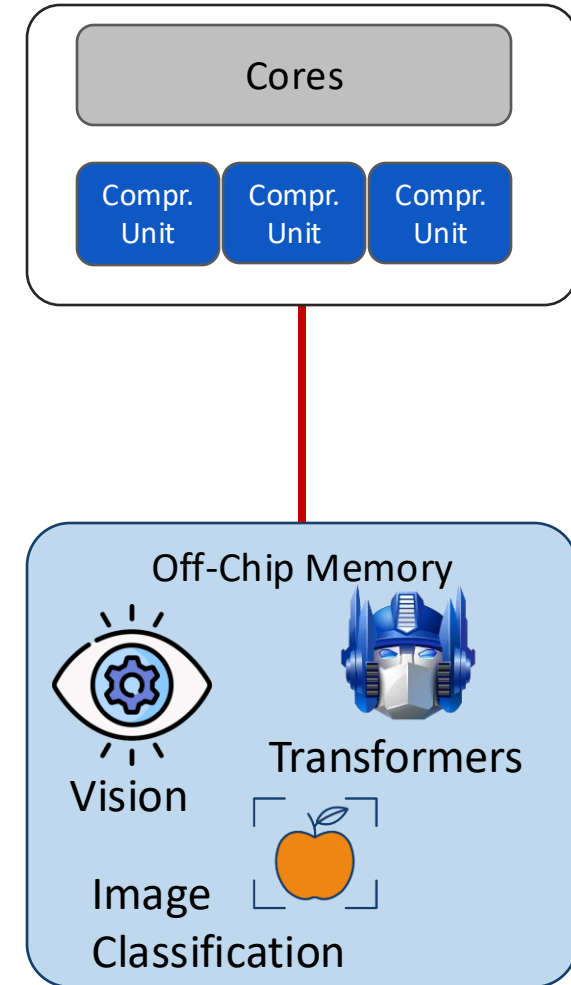
Atalanta: Brief Overview

- Frequent values need less than 1 bit
 - Arithmetic Coding [Langdon, 1984]
- Low Hardware Costs
- Programmer Transparent
- Lossless
- Inference + Training (forward pass)
- High Compression Rate



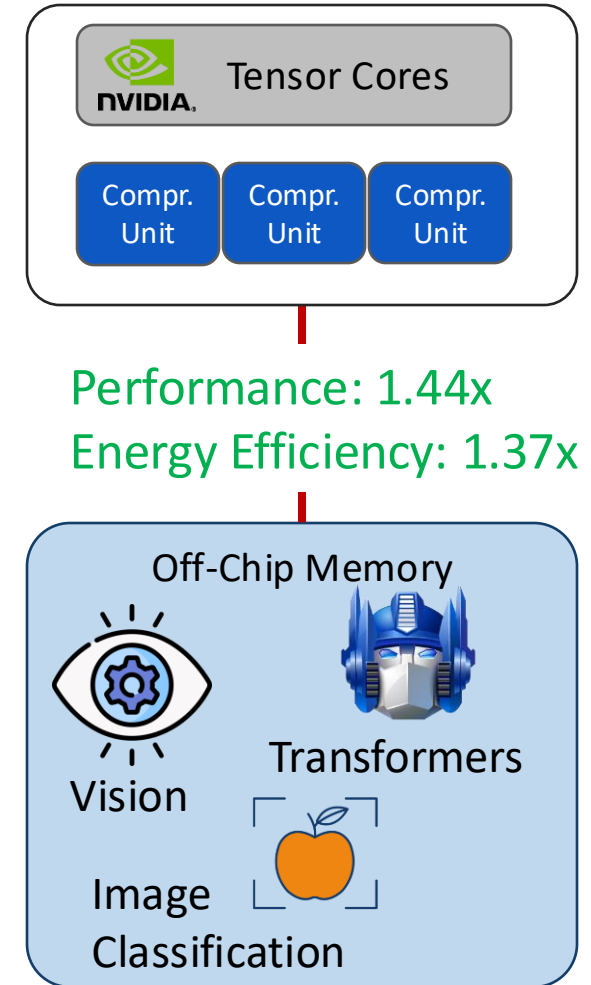
Atalanta: Brief Overview

- Frequent values need less than 1 bit
 - Arithmetic Coding [Langdon, 1984]
- Low Hardware Costs
- Programmer Transparent
- Lossless
- Inference + Training (forward pass)
- High Compression Rate
- High Energy Efficiency



Atalanta: Brief Overview

- Frequent values need less than 1 bit
 - Arithmetic Coding [Langdon, 1984]
- Low Hardware Costs
- Programmer Transparent
- Lossless
- Inference + Training (forward pass)
- High Compression Rate
- High Energy Efficiency
- **Seamlessly** Integrated with **AI Accelerators**



Atalanta: A Bit is Worth a “Thousand” Tensor Values

Alberto Delmas Lascorz, Mostafa Mahmoud, Ali Hadi Zadeh,
Miloš Nikolić, Kareem Ibrahim, [Christina Giannoula](#),
Ameer Abdelhadi, Andreas Moshovos

