Interferences between Communications and Computations in Distributed HPC Systems

Philippe Swartvagher

Interferences between Communications and Computations in Distributed HPC Systems

Philippe SWARTVAGHER - Alexandre DENIS, Emmanuel JEANNOT

TADaaM - INRIA, LaBRI, France

Distributed HPC systems

- HPC: *High-Performance Computing*
- Distributed: many computers connected through a network
- Common usecases: scientific computations, simulations (physics, chemistery, weather, astronomy, ...)
- Goal: get the highest computing performance
- One HPC cluster is composed of:
  - Several nodes...
  - ... composed of several computing units (CPUs, GPUs, ...)...
  - ... composed of many cores.

Computations and communications in parallel

- While many cores are computing, a core makes communications with other nodes
- A technic known for improving performances

Interferences between communications and computations ?

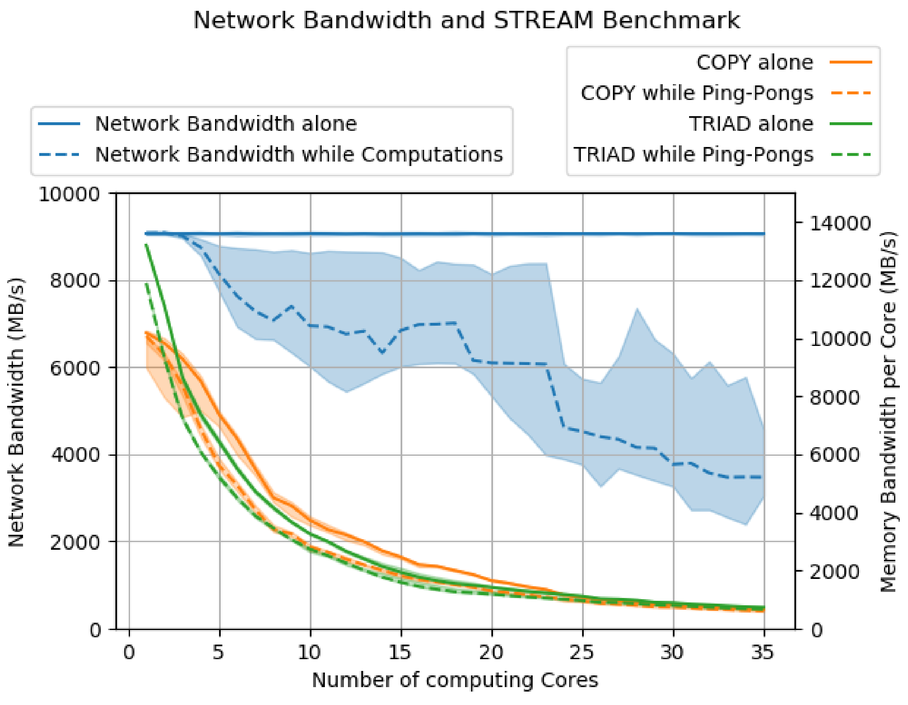
- Side-by-side communications and computations share common resources (memory bus)
- Impact of memory contention on computations already observed in the litterature
- **Can simultaneous communications and computations have a negative impact on each other ?**
  - Is there an impact on communications ?
  - Which factors change the impact of interferences ?

Methodology

- Goal: compare performances of communications and computations executed alone or simultaneously
- Own benchmark with the following steps:
  1. Computations alone
  2. Communications alone
  3. Computations with communications in parallel
- Parallelization with OpenMP, communications with MPI

Impact of memory contention

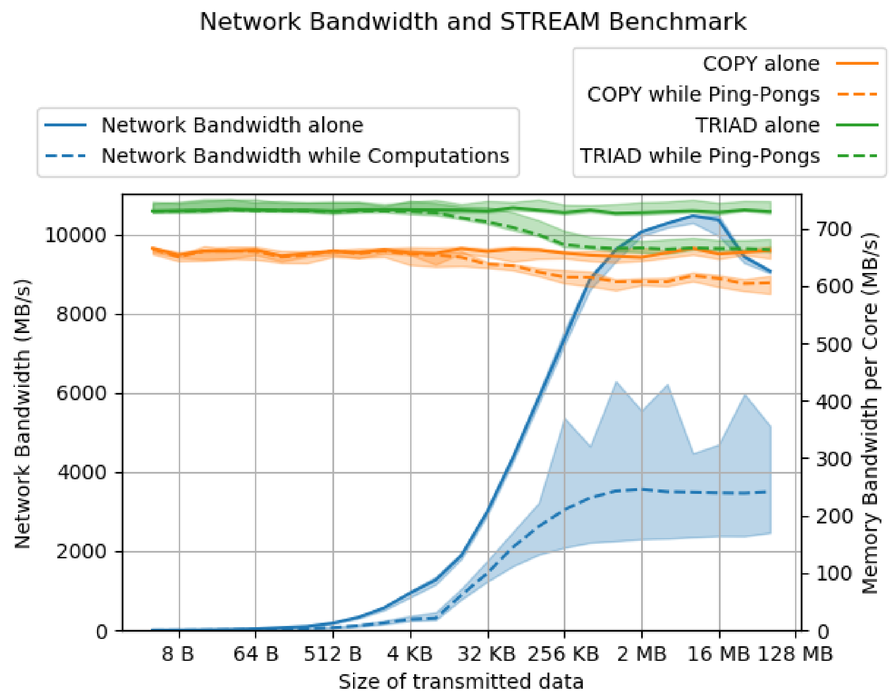
- Extreme case: computing cores executing memory-bound kernels



- The higher the number of computing cores, the higher the memory contention and the more communication performances lower

Impact of the size of transmitted data

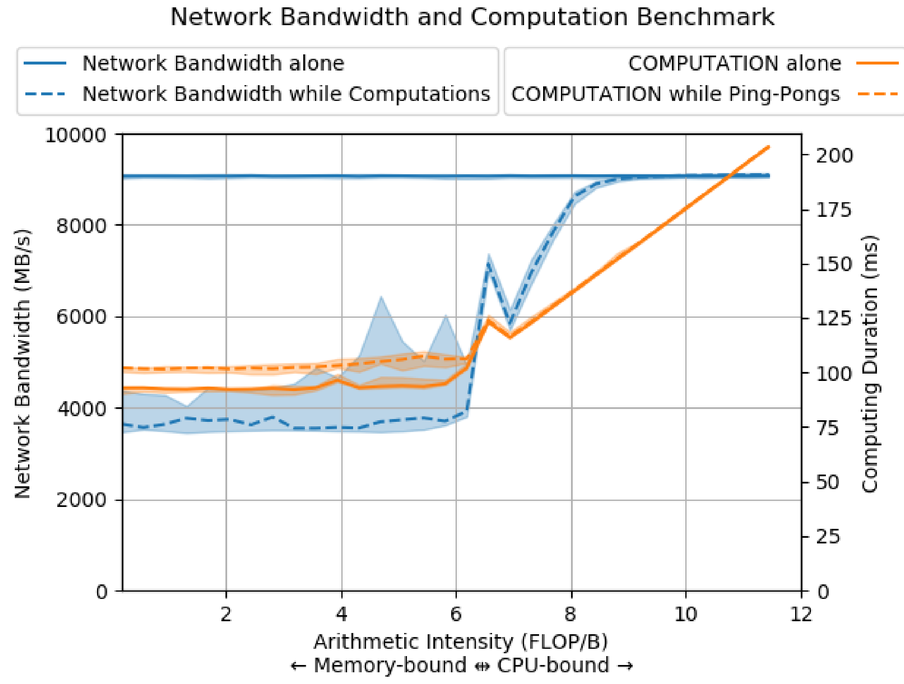
- Does the size of transmitted data across the network affect the memory contention ?



- The bigger the communicated data, the more it disturbs computations, and the more communications are disturbed by computations

Impact of arithmetic intensity

- Arithmetic intensity: number of arithmetic operations per accessed byte of memory
- Computing cores execute instructions with tunable arithmetic intensity



- Interferences disappear when computations become CPU-bound

Conclusion

- Computations can slow down communications
  - it depends on arithmetic intensity of computations
  - it depends on size of transmitted data
  - *vice-versa*, communications can impact computations too
- Future works:
  - Better understand origins of interferences to model and predict them
  - Consider data movements between main memory and GPUs
  - Take into account these interferences in scheduling made by runtime systems