Computational Science on Many-Core Architectures

360.252

Josef Weinbub
Karl Rupp

Institute for Microelectronics, TU Wien
http://www.iue.tuwien.ac.at/

Zoom Channel 621 2711 2607
Wednesday, November 22, 2023

Performance Modeling

Latency

- Bottleneck in strong scaling limit
- Ultimate limit for time stepping

Latency - Sources

- Network latency (Ethernet $\sim 20\mu$s, Infiniband $\sim 5\mu$s)
- PCI-Express latency (Kernel launches, $\sim 10\mu$s)
- Thread synchronization (barriers, locks, $\sim 1 - 10\mu$s)
- Memory latency (~ 100ns)

Performance Modeling

Load Imbalance

- Total execution time determined by slowest thread
- Focus on making the slowest thread fast
- Easy for static data structures (e.g. dense matrices)
- Hard for dynamic data structures (e.g. sparse matrices)

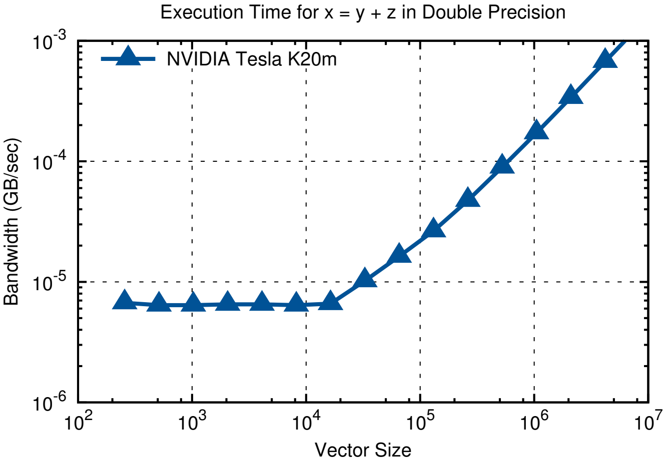
Amdahl's Law

- Total execution time T_{total} given by time spent in serial and parallel parts
- $T_{\text{total}} = T_{\text{serial}} + T_{\text{parallel}} / \#\text{processors}$
- Speed-up limited by serial portion of an algorithm

Performance Modeling: Vector Addition

Vector Addition

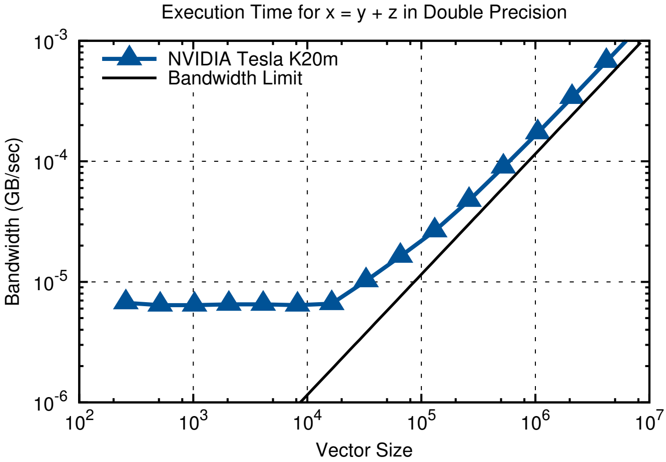
- $x = y + z$ with N elements each
- 1 FLOP per 24 byte in double precision
- Limited by memory bandwidth $\Rightarrow T_2(N) \overset{?}{\approx} 3 \times 8 \times N/\text{Bandwidth} + \text{Latency}$



Performance Modeling: Vector Addition

Vector Addition

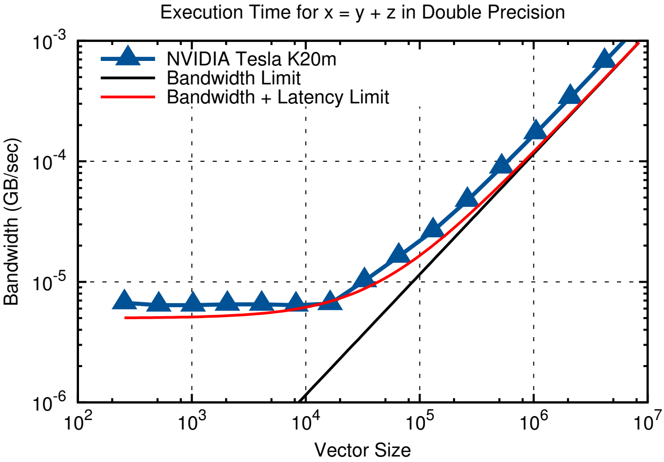
- $x = y + z$ with N elements each
- 1 FLOP per 24 byte in double precision
- Limited by memory bandwidth $\Rightarrow T_2(N) \overset{?}{\approx} 3 \times 8 \times N/\text{Bandwidth} + \text{Latency}$



Performance Modeling: Vector Addition

Vector Addition

- $x = y + z$ with N elements each
- 1 FLOP per 24 byte in double precision
- Limited by memory bandwidth $\Rightarrow T_2(N) \overset{?}{\approx} 3 \times 8 \times N/\text{Bandwidth} + \text{Latency}$



Performance Modeling: Vector Addition

Vector Addition

- $x = y + z$ with N elements each
- 1 FLOP per 24 byte in double precision
- Limited by memory bandwidth $\Rightarrow T_2(N) \overset{?}{\approx} 3 \times 8 \times N/\mathrm{Bandwidth} + \text{Latency}$