30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)
Lyon, France, September 24-27, 2018

Predicting the Performance Impact of Increasing Memory Bandwidth for Scientific Workflows

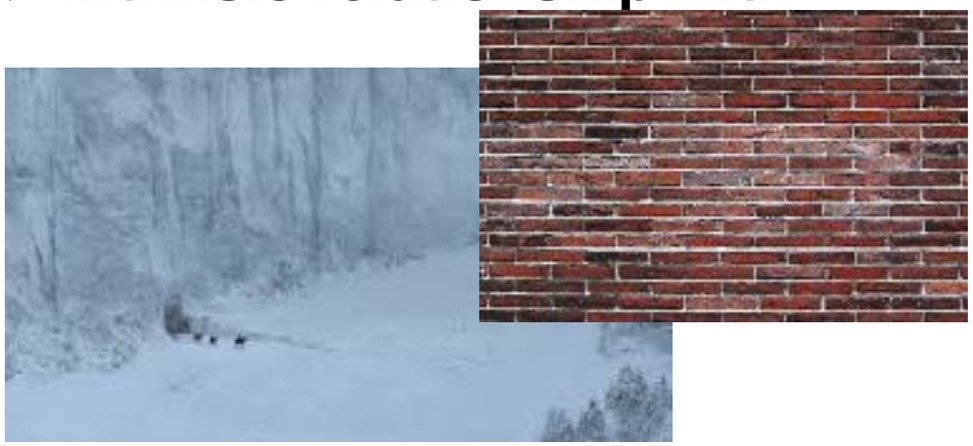
Nelson M. Gonzalez, Jose Brunheroto, Fausto Artico, Yoonho Park
IBM T. J. Watson Research Center, New York, USA

Tereza Carvalho
Escola Politécnica, University of São Paulo, Brazil

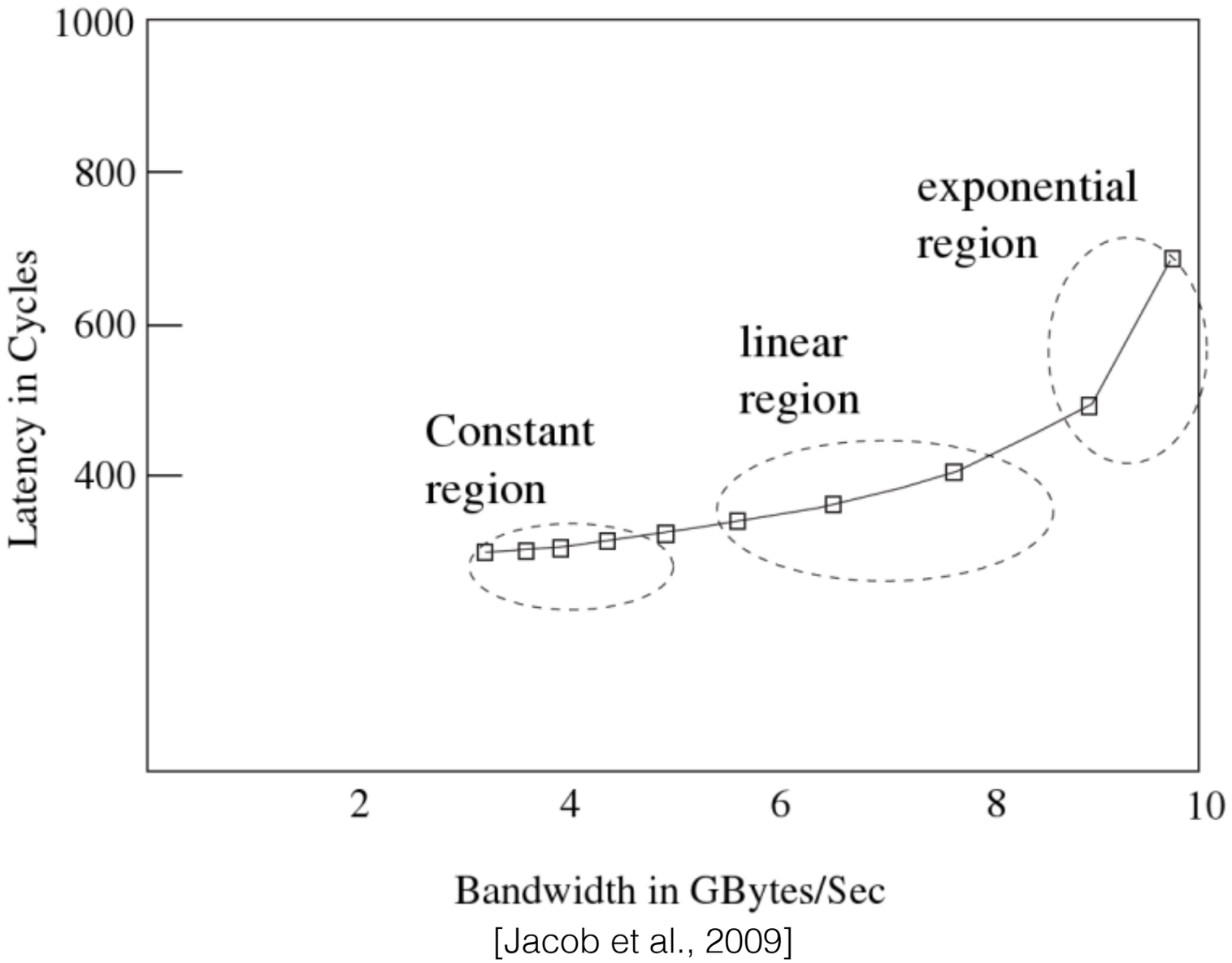
Charles C. Miers, Maurício A. Pillon, Guilherme P. Koslovski
Graduate Program in Applied Computing — Santa Catarina State University — Joinville — Brazil

Context: Memory wall

- A well-known problem
  - *"The first hardware-related [design] issue is memory bandwidth: the benchmarks suggest that it is not keeping up with CPU speed […]. If memory does not improve dramatically in future machines, some classes of application may be limited by memory performance"* [J. Ousterhout, 1990]
  - *"chips are largely able to execute code faster than we can feed them with instructions and data"* [R. Sites, 1996]
  - **Memory wall: the performance is completely dependent upon memory speed**
    - **Initially defined in terms of latency —> intrinsic relationship with bandwidth [Radulovic et al., 2015]**



Context: Latency vs. Bandwidth curve



[Jacob et al., 2009]

Maximum sustained bandwidth = 10 GB/s

Context: Memory wall

- We have new hardware!
  - Several technologies such as Hybrid Memory Cube (HMC), Rambus Direct DRAM, Double Data Rate (DDR), and High Bandwidth Memory (HBM)
  - For instance: HMC should be able to deliver up to 15x the bandwidth of regular DRAM (~480 GB/s)

- The benefit of adopting these technologies depend on the target application [Radulovic et al., 2015]

- **The performance effects of improving bandwidth still lack exploration**
  [Wang et al., 2016, Wang et al., 2014]

Our goal

- **Investigation of the effects of improving memory bandwidth to scientific workflows**
  - **Define a performance model to predict the performance for a particular application**
  - **Investigate the correlation between the bandwidth used by an application and how efficiently an increase to total bandwidth is converted into performance**

Outline

- Memory bandwidth model
  - Curve η x *F*
  - Discretization and iterative method
- Experimental methodology
  - Platform & bandwidth variation
  - Applications & tools
- Analysis
- Conclusion & future work

Memory Bandwidth Model: Curve η x *F*

- **Memory-bound applications benefit more from an increase in total available bandwidth than application that are not memory bound**

B = effective memory bandwidth used by application

S = total sustained bandwidth

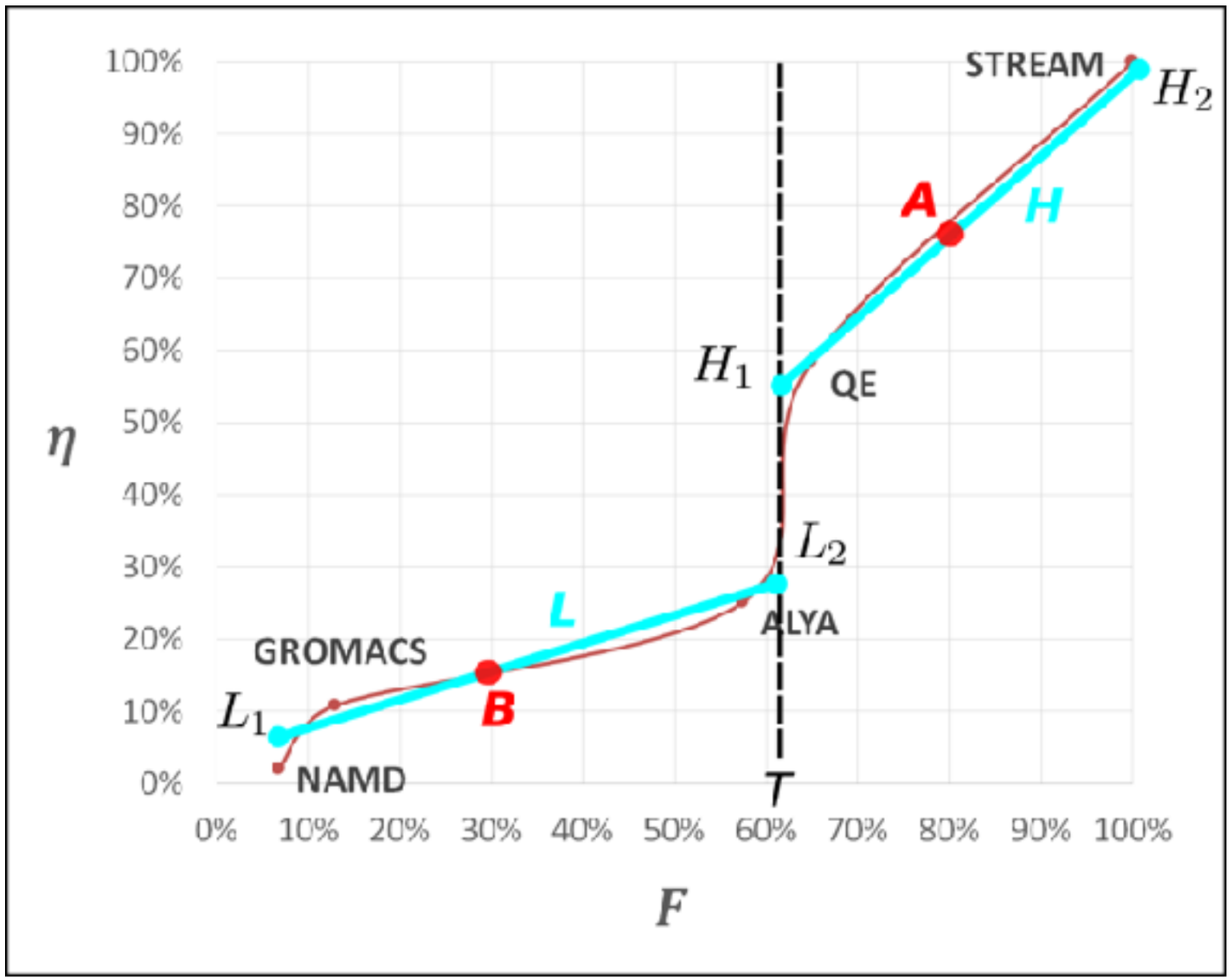
F = fraction of sustained bandwidth used by application —> *B/S*

I_B = increase in effective bandwidth

I_S = increase in total sustained bandwidth

η = efficiency metric —> I_B / I_S

Memory Bandwidth Model: Curve η x *F*



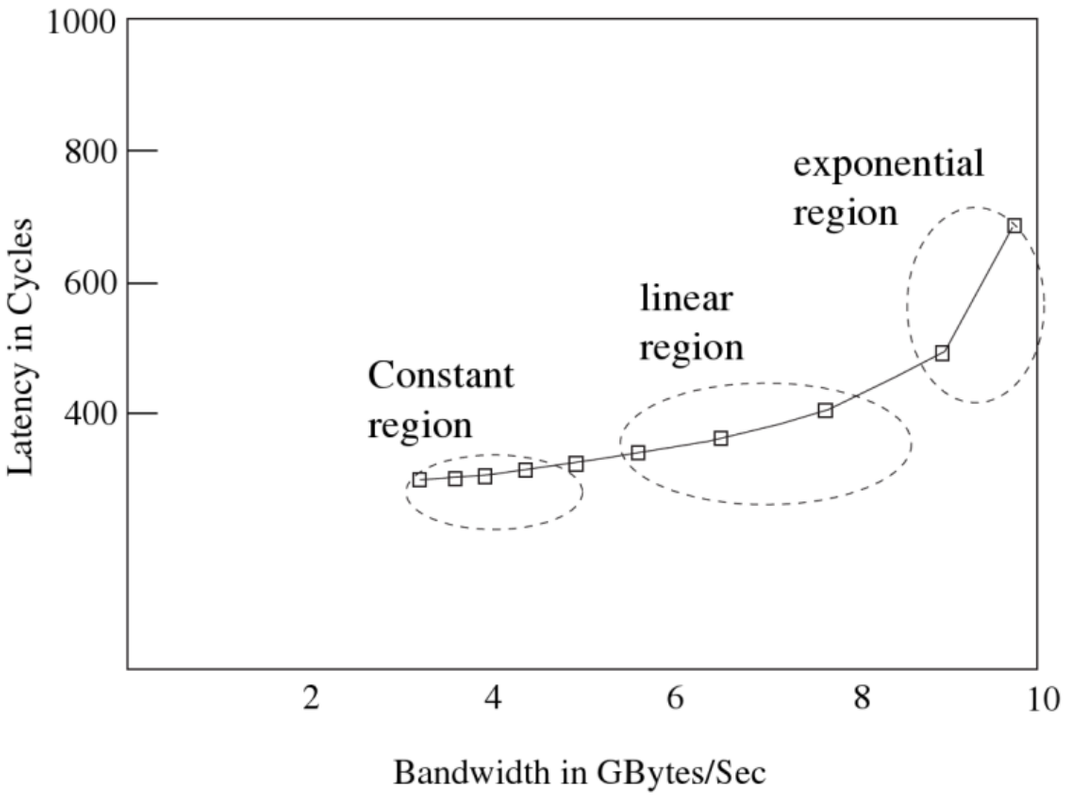
Applications and data from [Radulovic et al., 2015]

- **Applications with higher *F* demonstrate a better efficiency to convert bandwidth increase into application performance**

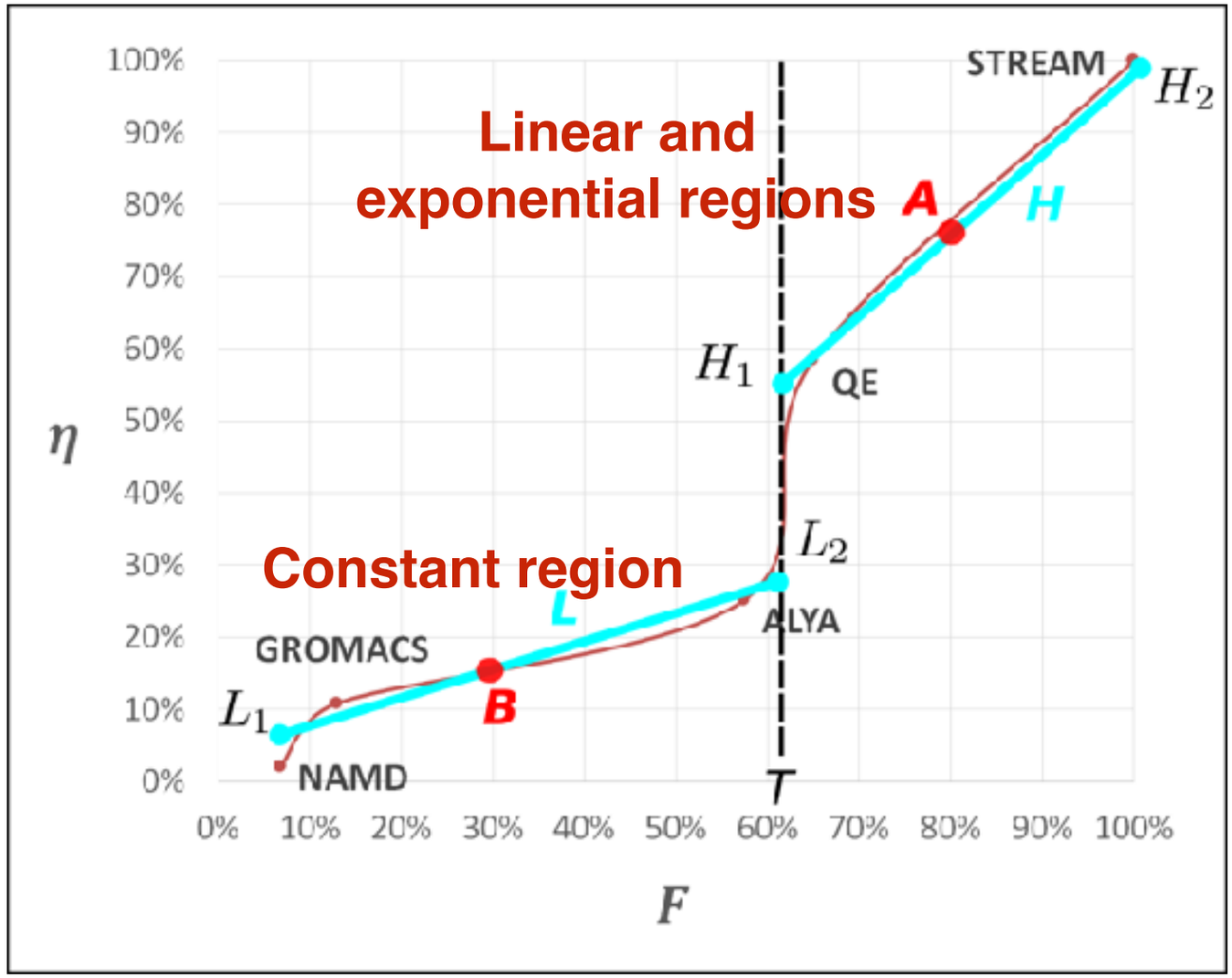
- Regions *H* and *L* are separated by an abrupt slope (*T*)

- ***A*** = memory-bound applications
- ***B*** = no memory intensive

Memory Bandwidth Model: Regions



[Jacob et al., 2009]



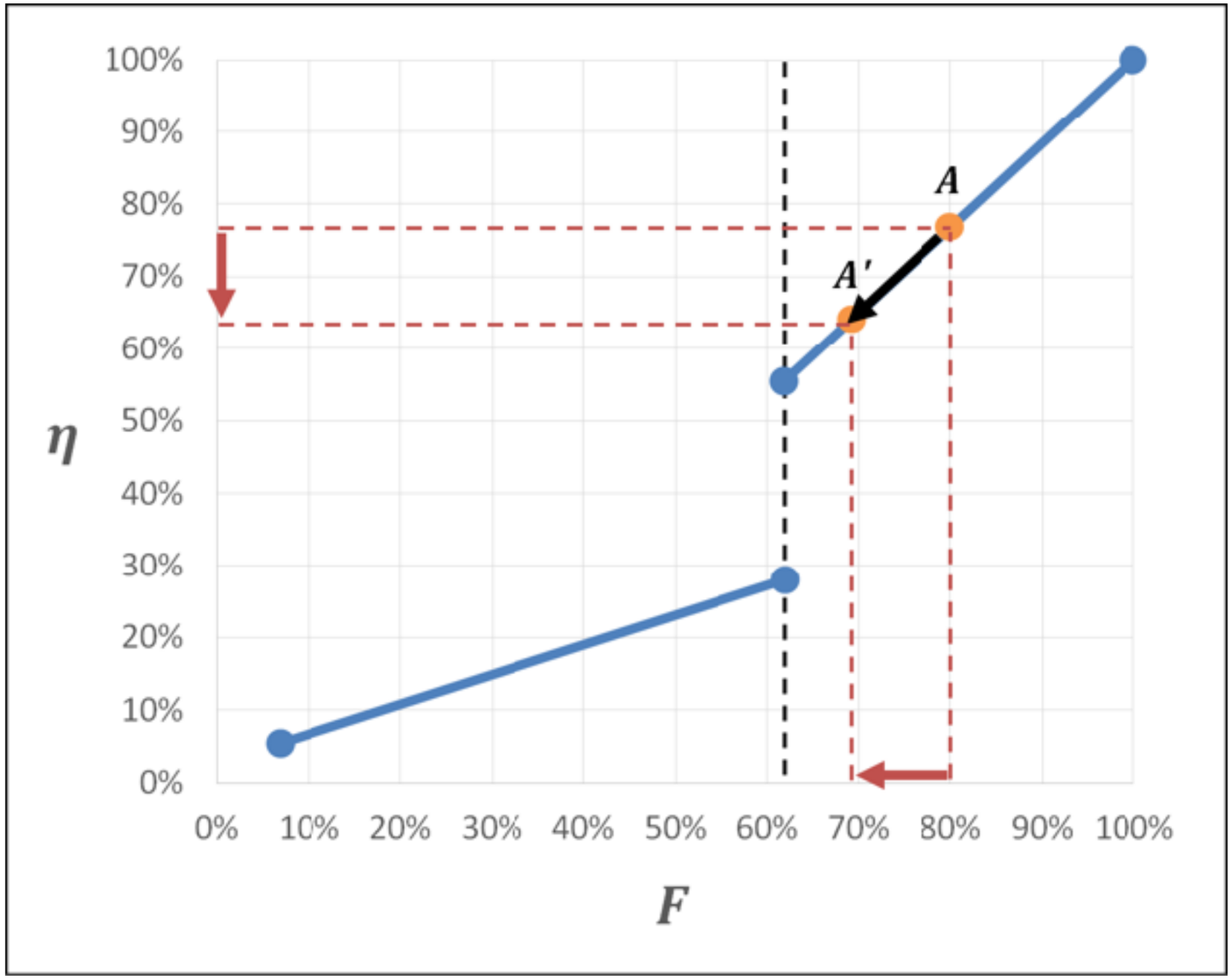
Memory Bandwidth Model: Discretization and Iterative Method

- **A set of expressions and a simple algorithm were proposed to iteratively compute *B***

- Key points:
  - We must compute with *k* increments of α instead a single one

  - Granularity matters: scheduling, cloud reservations, hardware acquisition

  - Having more bandwidth do not improve application performance (*F = B/S*)



Outline

- Memory bandwidth model
  - Curve η x *F*
  - Discretization and iterative method
- **Experimental methodology**
  - **Platform & bandwidth variation**
  - **Applications & tools**
- Analysis
- Conclusion & future work

Experimental methodology: Applications and platform

- The model was built based on the experimental results from specialised literature [Radulovic et al., 2015]

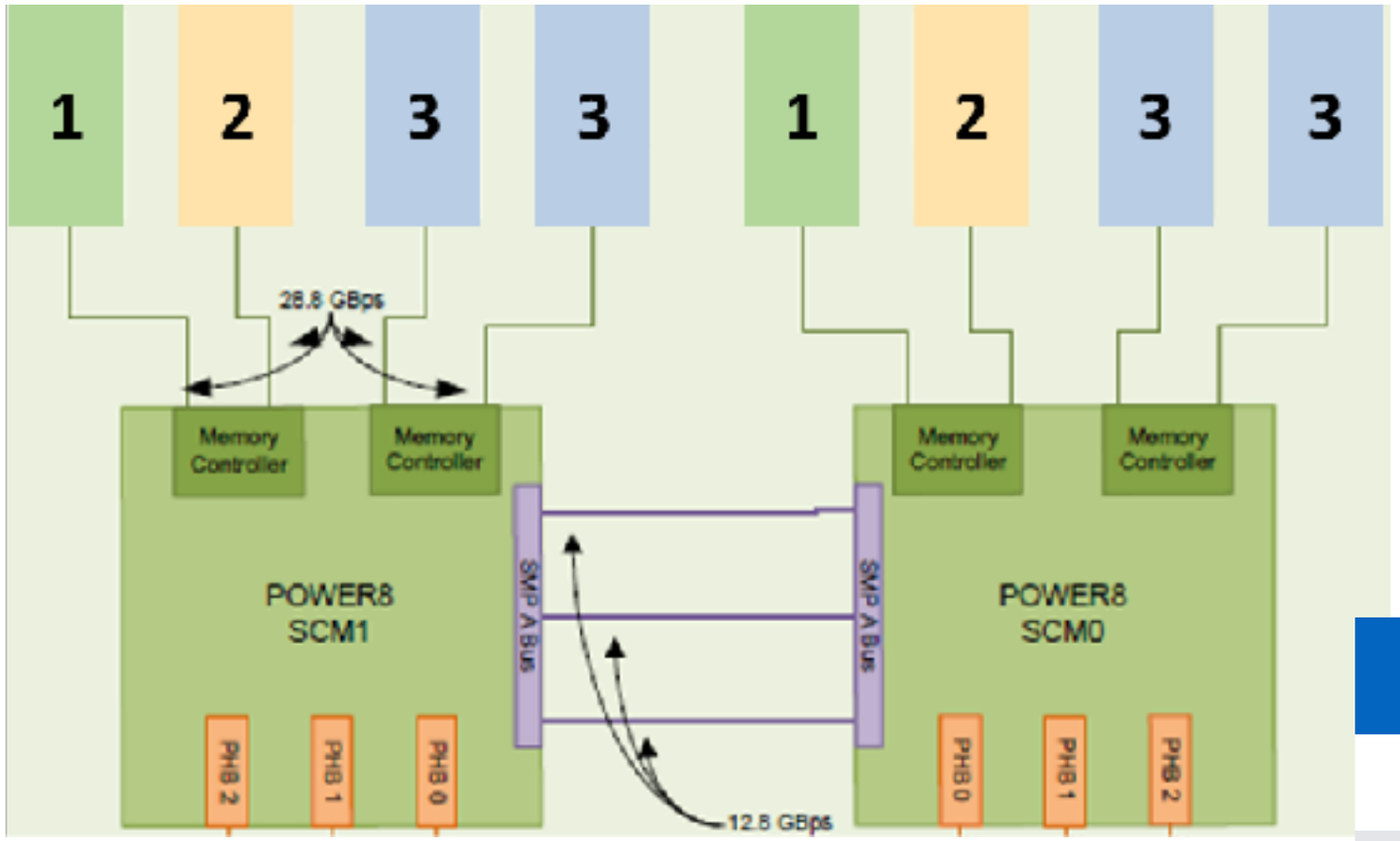
- **We analyze the model and assumptions executing on different hardware platform and applications**
  - Bandwidth variation is based on hardware configuration instead of clocks

- **4 applications**
  - **STREAM: memory benchmark**
  - **GTC-P, HPCG and MILC: applications and benchmarks from Crossroads/NERSC-9 APEX**
  - **MPI-based applications (20 processes)**

- Platform
  - IBM Power System S822LC model 8335-GTA (Firestone): 16 nodes (only 1 was used). Each node contains two POWER8 sockets with 10 cores.
  - 512 GB of main memory divided in 8 raises.

Experimental methodology: Bandwidth variation

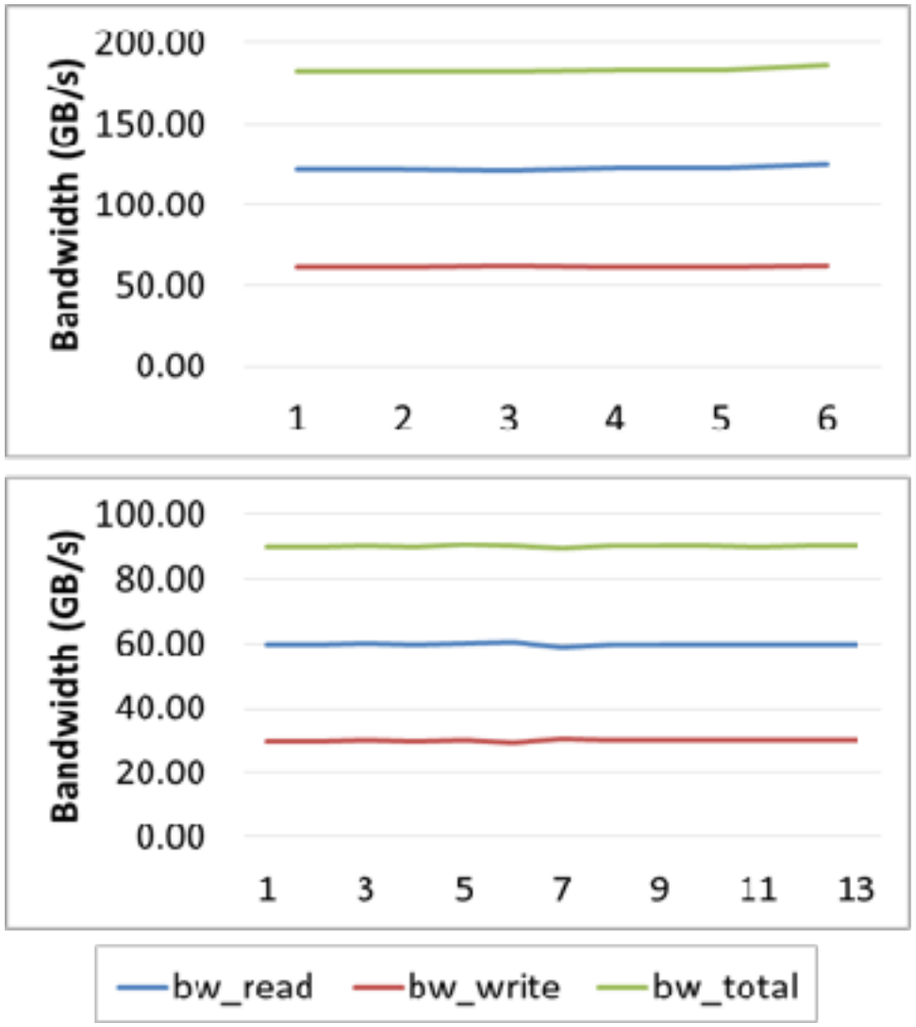
- 3 groups. Experimental results correspond to the transition from configuration II to I (half to full bandwidth)



	Configuration	Total peak	Relative
I	#1 + #2 + #3	230.4 GB/s	100%
II	#1 + #2	115.2 GB/s	50%
III	#1	57.6 GB/s	25%

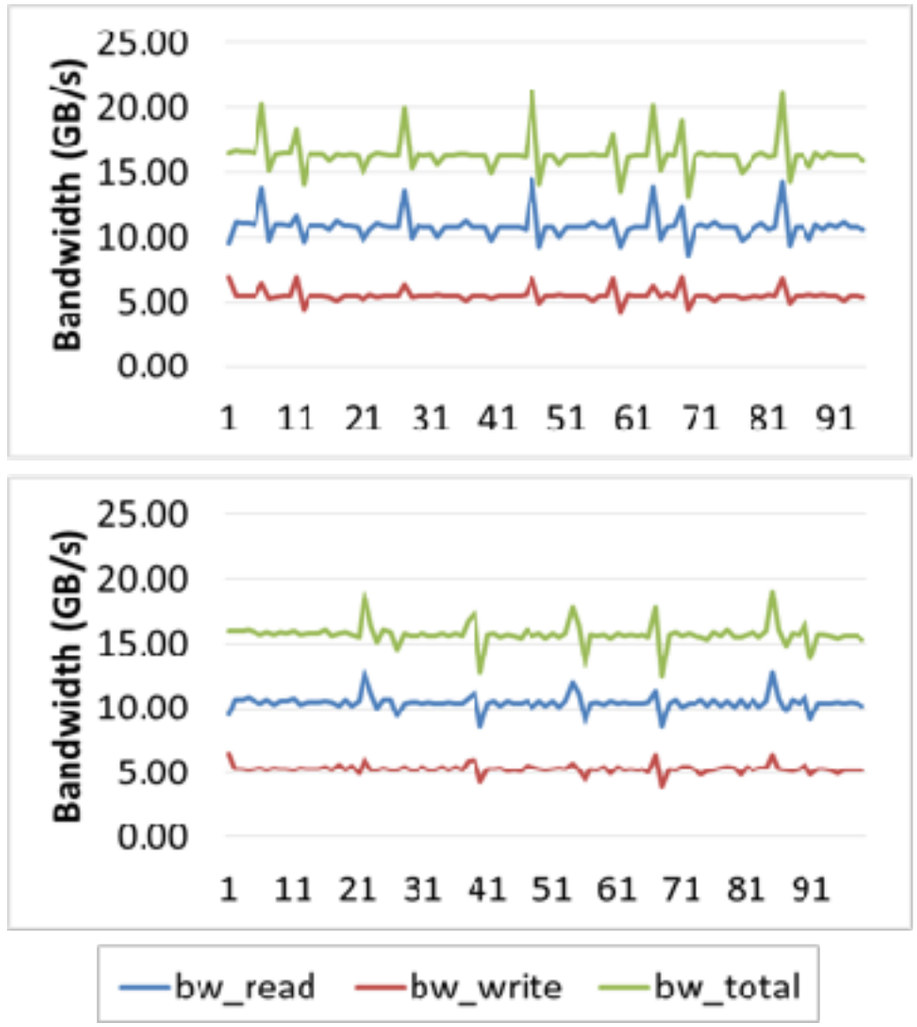
Analysis: Bandwidth (read, write and total)

STREAM benchmark



double bw

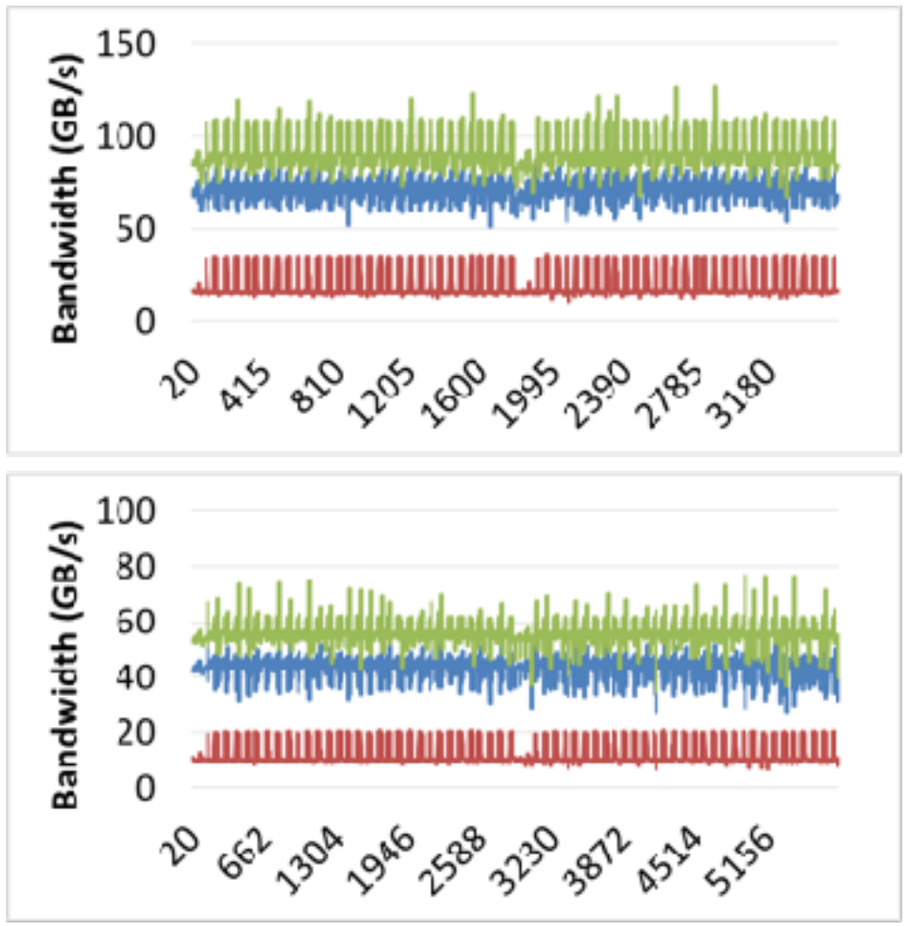
GTC-P



double bw

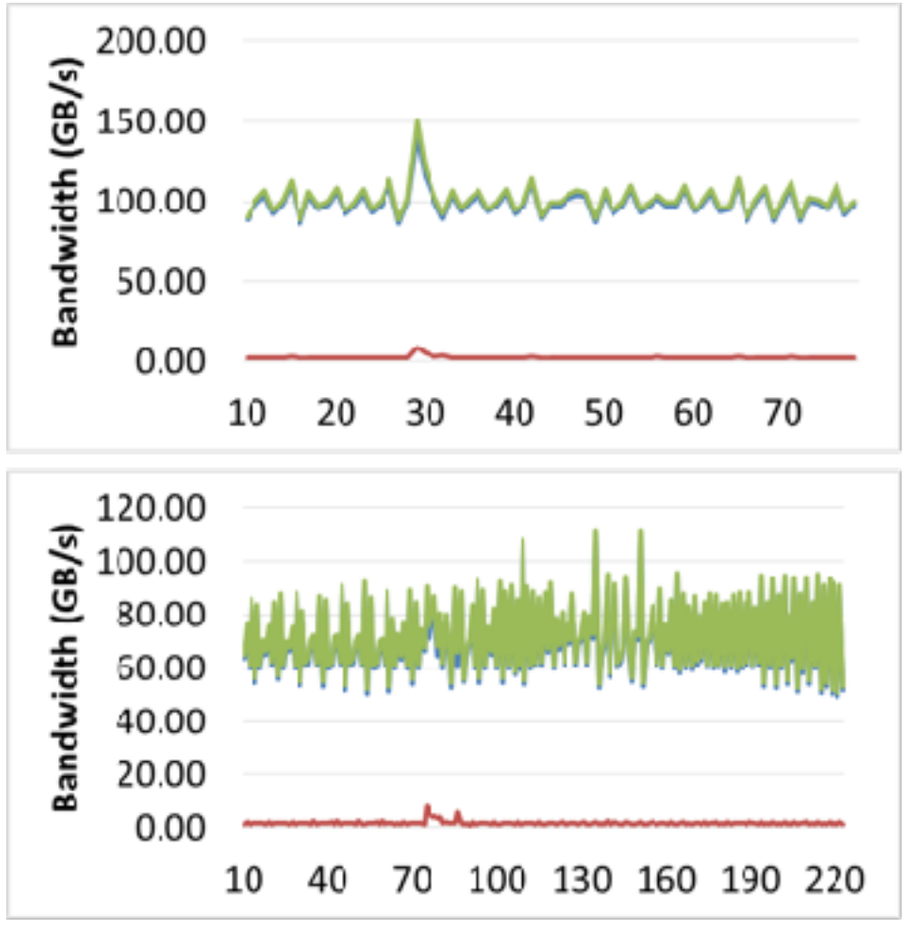
Analysis: Bandwidth (read, write and total)

MILC



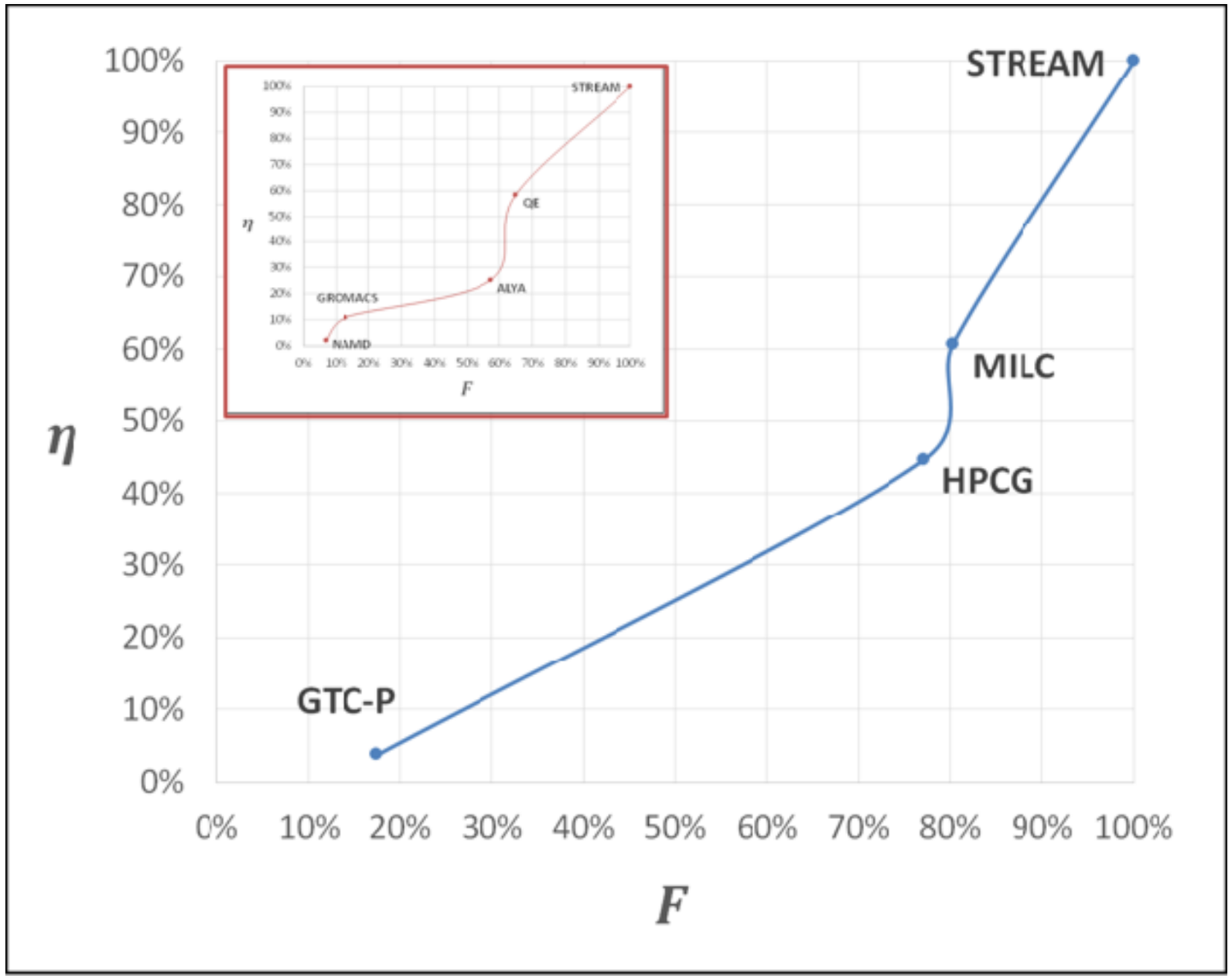
double bw

HPCG



double bw

Analysis: Curve η x *F*



	B (GB/s)	*F*	*B'* (GB/s)
STREAM	89.77	100.0%	183.84
MILC	56.25	80.29%	91.52
HPCG	69.30	77.19%	101.75
GTC-P	15.74	17.53%	16.36

	F'	I_B	η
STREAM	100.0%	104.79%	100.0%
MILC	64.29%	62.68%	60.76%
HPCG	55.53%	46.83%	44.69%
GTC-P	8.90%	3.92%	3.74%

Conclusion & perspectives

- Investigation on the effects of memory bandwidth on application performance (scientific workflow benchmarks)

- **Performance model to understand memory bound regions of applications**

- **Observations:**
  - **Memory bound applications benefit more from an improvement to total available bandwidth**
  - **The performance gains of improving available bandwidth gradually diminish as this bandwidth increases**

- **Limits:**
  - **The performance behaviour might depend on hardware and software properties**

- Future work
  - Different hardware and software platforms
  - Virtualized environments (IaaS clouds)
  - Multithreading

30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)
Lyon, France, September 24-27, 2018

Thank you!

Predicting the Performance Impact of Increasing Memory Bandwidth for Scientific Workflows

Nelson M. Gonzalez, Jose Brunheroto, Fausto Artico, Yoonho Park
IBM T. J. Watson Research Center, New York, USA
Tereza Carvalho
Escola Politécnica, University of São Paulo, Brazil
Charles C. Miers, Maurício A. Pillon, Guilherme P. Koslovski
Graduate Program in Applied Computing — Santa Catarina State University — Joinville — Brazil