Delay Flow Mechanisms on Clusters

Ayesha Afzal, Georg Hager, Gerhard Wellein
{ayesha.afzal, georg.hager, gerhard.wellein}@fau.de

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

High Performance Computing | Erlangen Regional Computing Center | Friedrich-Alexander University Erlangen-Nürnberg, Germany

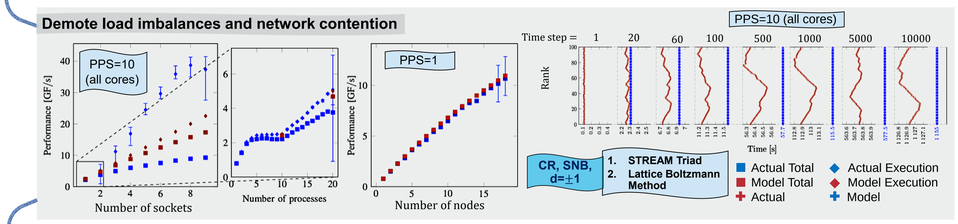
Problem: Analytic Performance Model for Distributed-Memory Applications

Distributed Runtime Model — Lock-step, Syncronicity, Assumption Failure ?

≠ **Execution Models**: ECM (L1 Fastest, L2 Fast, L3 Less Fast, MEM Slow), Roofline (Compute-bound, Memory-bound, P [F/s], I [F/B])

+ **Communication Models**: LogP Hockney (BW [B/s], Msg [B], L [s], BW)

Demote load imbalances and network contention



Influence Factors

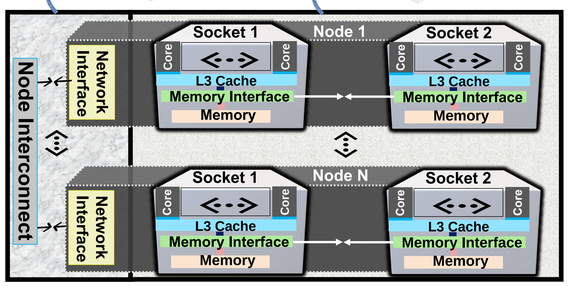
Application: Computation and Communication Characteristics (Nomenclature)

Hierarchical Topology (affinity matters): **Node-level**: ccNUMA SMP nodes with multi-socket, multi-core CPUs of Hyperthreads and SIMD cores. **Network-level**: multi-routes of different characteristics (like latency and asymptotic bandwidth), multi-ways of handling interconnet protocols for intra-chip, inter-chip/-node.

Hierarchical Parallel Programming Models

Outer-level Across Node Model: MPI

Inner-level Inside Node Model:



1. **MPI:** Process Parallelism per SHMEM Node/Socket/Core — Communication (Intra-chip + Inter-chip)

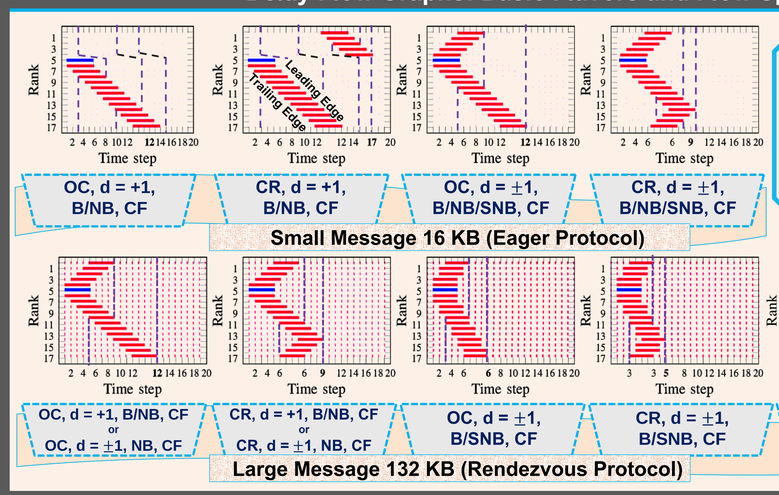
2. **X (e.g., OpenMP):** Thread Parallelism per Core/HT/Vectorization — Communication X Folk-Join Overhead ✓

3. **MPI + X (e.g., OpenMP):** Process and Thread Parallelism — Demote Communication + Folk-Join Overhead

Motivation: Model the Mystery of Non-synchronism

Distributed Runtime Model = **Execution Runtime Models** (Roofline ○ ECM) + **Communication Runtime Models** (Hockney ○ LogP) + **Non-synchronicity Mechanism** (Accelerator ○ Slowdown)

Delay Flow Graphs: Basic Flavors and Flow Speed on Silent Systems



Communication Topologies

Environmental Setup to Demote

- **Execution Delays** (Compute-bound vdivpd, fixed cy/inst., no data transfer boundness)
- **System Noise** (Single, fully nonblocking IB/OP leaf switch)
- **Communication Delays** (PPN=1 → no intranode comm. full non-blocking bandwidth with small Inter-node comm. overhead → no newtwork contention)

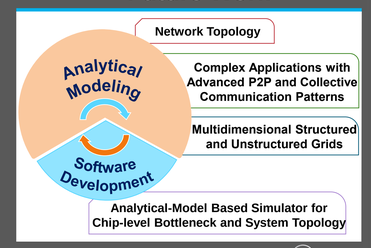
Propagation Speed

$$v_{silent} = \frac{\sigma.d}{T_{exec} + T_{comm}} \left[\frac{ranks}{s}\right]$$

$$\sigma: \begin{cases} 2, & \text{if Bidirectional R, B/SNB mode} \\ 1, & \text{else} \end{cases}$$

T_{comm}: Runtime of comm. happening in delay flow
d: distance between communicating processes

Future Work



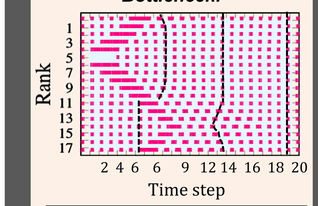
References

1. A. Afzal et al. Propagation and Decay of Injected One-Off Delays on Clusters: A Case Study. Accepted for IEEE Cluster 2019, Albuquerque, NM, September 23-26, 2019. Preprint: arXiv:1905.10603.
2. T. Hoefler et al. LogGOPSim - Simulating Large-Scale Applications in the LogGOPS Model. DOI: 10.1145/1851476.1851564.
3. S. Markidis et al. Idle waves in high-performance computing. DOI: 10.1103/PhysRevE.91.013306.
4. IB. Peng et al. Idle period propagation in message-passing applications. DOI: 10.1109/HPCC-SmartCity-DSS.2016.0134.
5. JD. McCalpin et al. Memory bandwidth and machine balance in current high performance computers. *IEEE computer society technical committee on computer architecture (TCCA) newsletter*, 2(19–25), 1995.

Global Non-synchronism Phenomenology: Noise-assisted Long-distance Correlations and Structure Formation

Communication Variations yields

Network Bandwidth Bottleneck:



CR, d=±1, R Mode, NB, C load

Exchange of Large Messages (3MB) to Multiple Neighbours

$$v_{silent} = \frac{\sigma.d}{T_{exec} + T_{comm}^{BW(PPS)}} \left[\frac{ranks}{s}\right]$$

Execution Variations yields

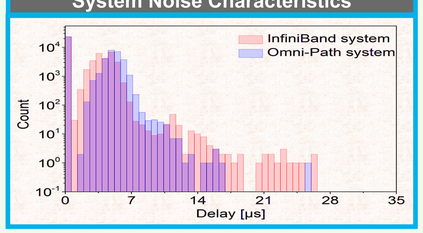
STREAM Triad
MPI Process Parallelism per Core
CR, d=±1, E Mode, NB, CF

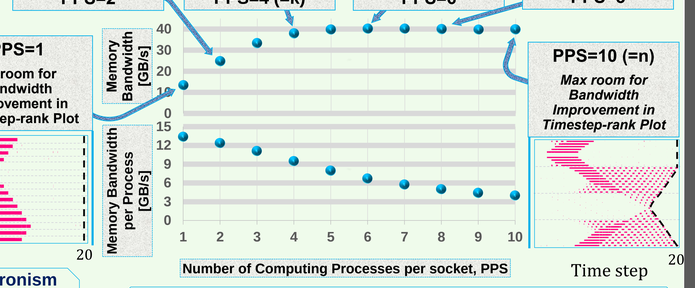
Memory Bandwidth Bottleneck:

$$v_{silent} = \frac{\sigma.d}{T_{exec}^{BW(PPS)} + T_{comm}} \left[\frac{ranks}{s}\right]$$

Variations: Random/ Periodic with System Topology/ Application Time step

System Noise Characteristics



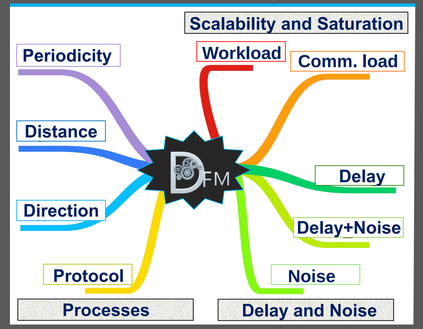


Synchronism (Chip MEM Saturation by 1 rank) <--> **Partial Non- synchronism** <--> **Perfect Non- synchronism**

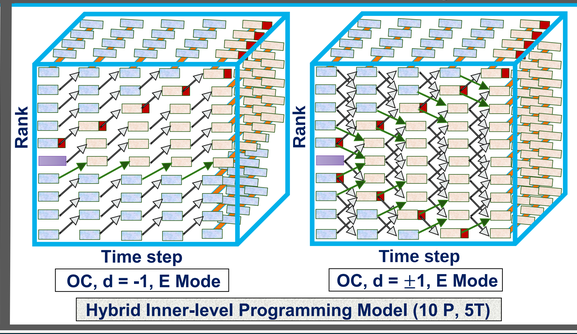
	Lock-step Synchronism	Non-synchronism
Execution-comm. Overlap	Perfect Non-overlap	Partial or Perfect Overlap
Memory Bus/Network Utilization	Intermittent	More Continuous
Processes in Saturated Resource(s) Regime	P = n	$P < n$
Resource(s) Bandwidth Per Process	Less (contented)	More/Full (less/no contented)
Execution time, $T_{exec}^{BW(PPS)}$	Slowest	Faster
Idle time, $T_{idle}^{BW(PPS)}$	Minimum	Lengthen
Idlewave Decay	Weak	Strong

Triad ($2\ MPI\ PPS * 5\ OpenMP\ T, V_{mem} = 4.8\ GB, Msg = 29\ MB, b_{IB} \approx 3\ GB/s$)

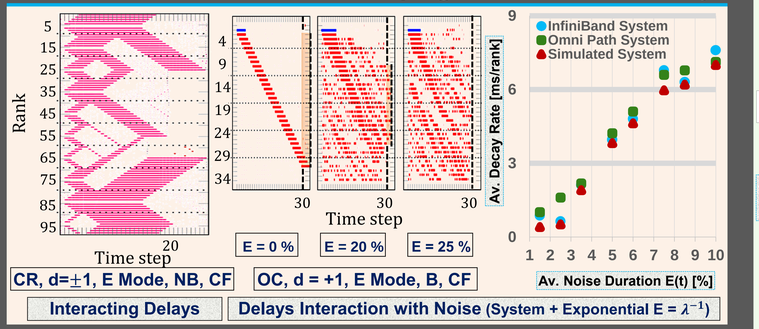
Parameter Space



Communication-assisted Delay Flow



Interaction & Noise-assisted Damping + Elimination of Delays



Workload	Periodicity for Boundary Conditions	Direction	Distance of Communicating Processes	Commuication Protocol	Communication Routine	Communication Load
: Compute Bound	OC : Open Chain	+ or - : Uni directional	d = 1 : Direct Neighbour Processes	E : Eager	B : Blocking (sendrecv, sendrecv,…)	CF : Contention Free
: Memory Bound	CR : Close Ring	± : Bi directional	d = 2, 3, … , n : Indirect Neighbour Processes	R : Rendezvous	NB : Non-blocking (isend, irecv, isend, irecv, …, waitall) SNB : NB + split wait (isend, irecv, waitall, isend, irecv, waitall, …)	C : Contention in Network

KONWIHR OMI4papps