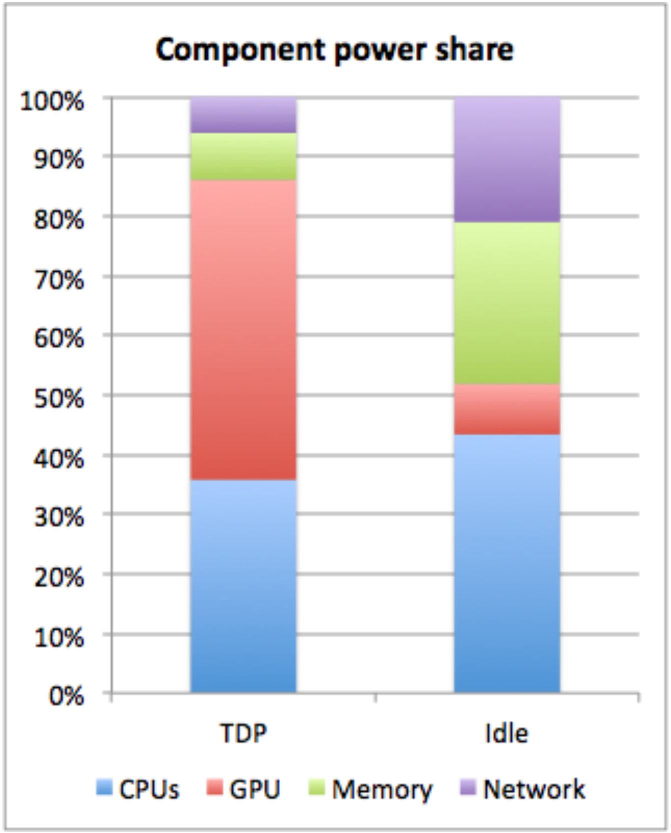
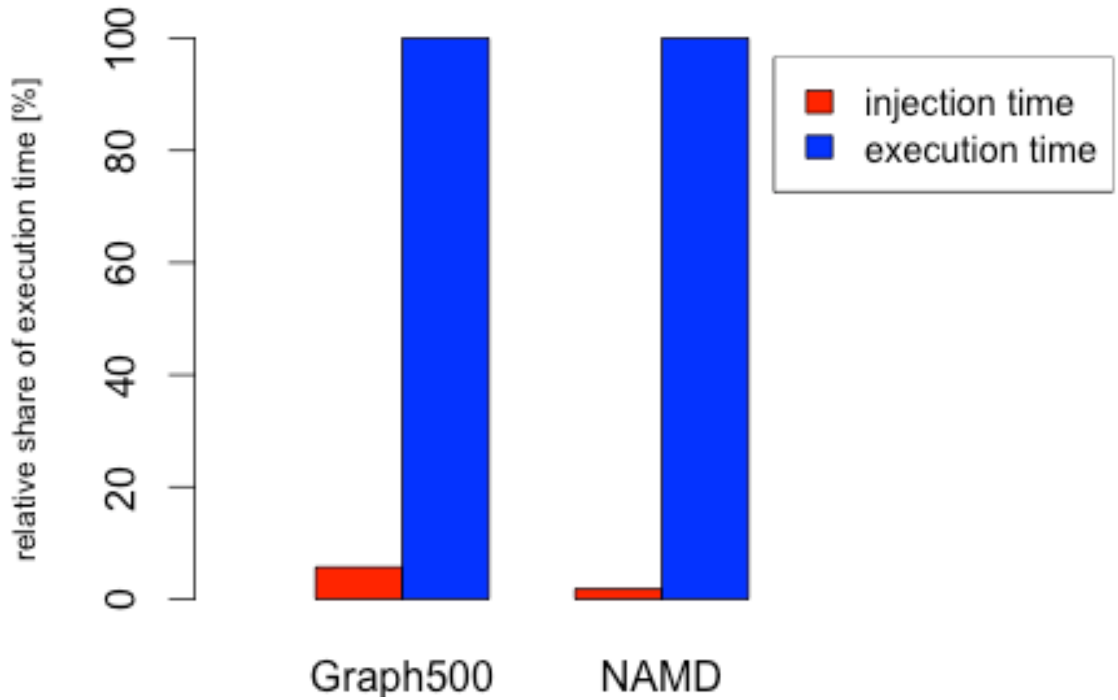
Motivation – Network Power

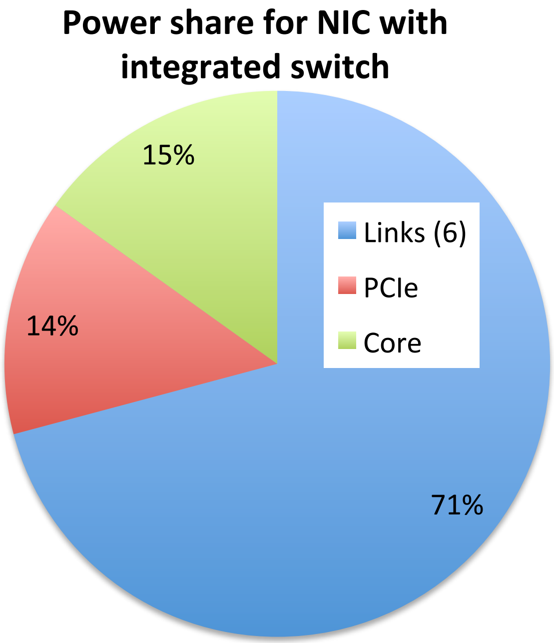


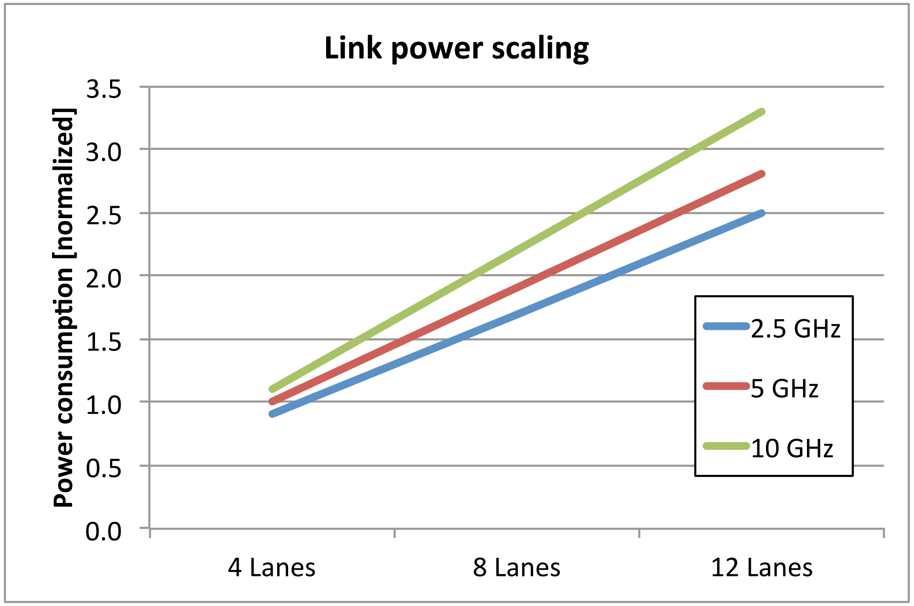


- Promising insights from first experiments:
  - At low utilization the network power share rise up to more than 20%
  - Network ports are idling the most time

RUPRECHT-KARLS-
UNIVERSITÄT
HEIDELBERG

Motivation – NIC Power Distribution





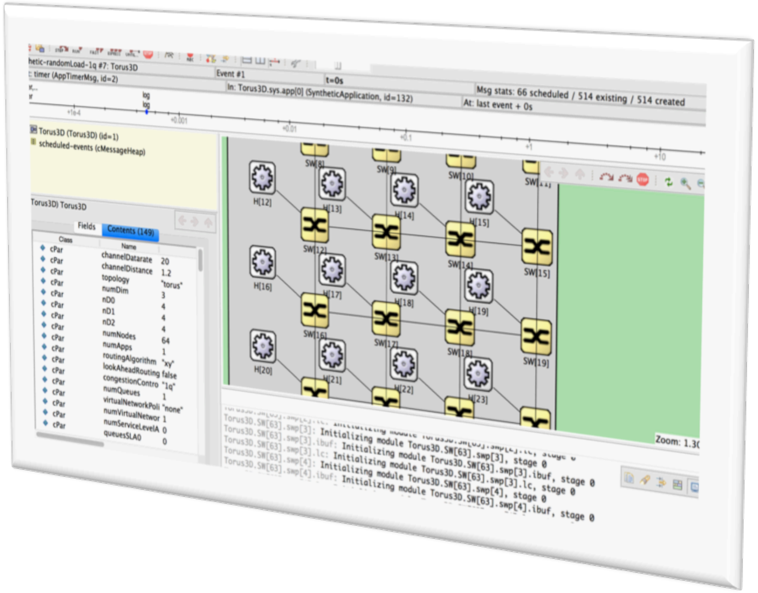
- Serialization technology dominates power consumption
  - Clock recovery, high frequency, equalization, pre-emphasis, …
- It is link width that matters, not frequency
  - CML = **Current** Mode Logic
  - Linear scaling for 10GHz case
  - Frequency dependent part is CMOS only

Concept

- How to trade on these insights?
  - Change link configurations according to current demands dynamically

    => Policy necessary for these decisions

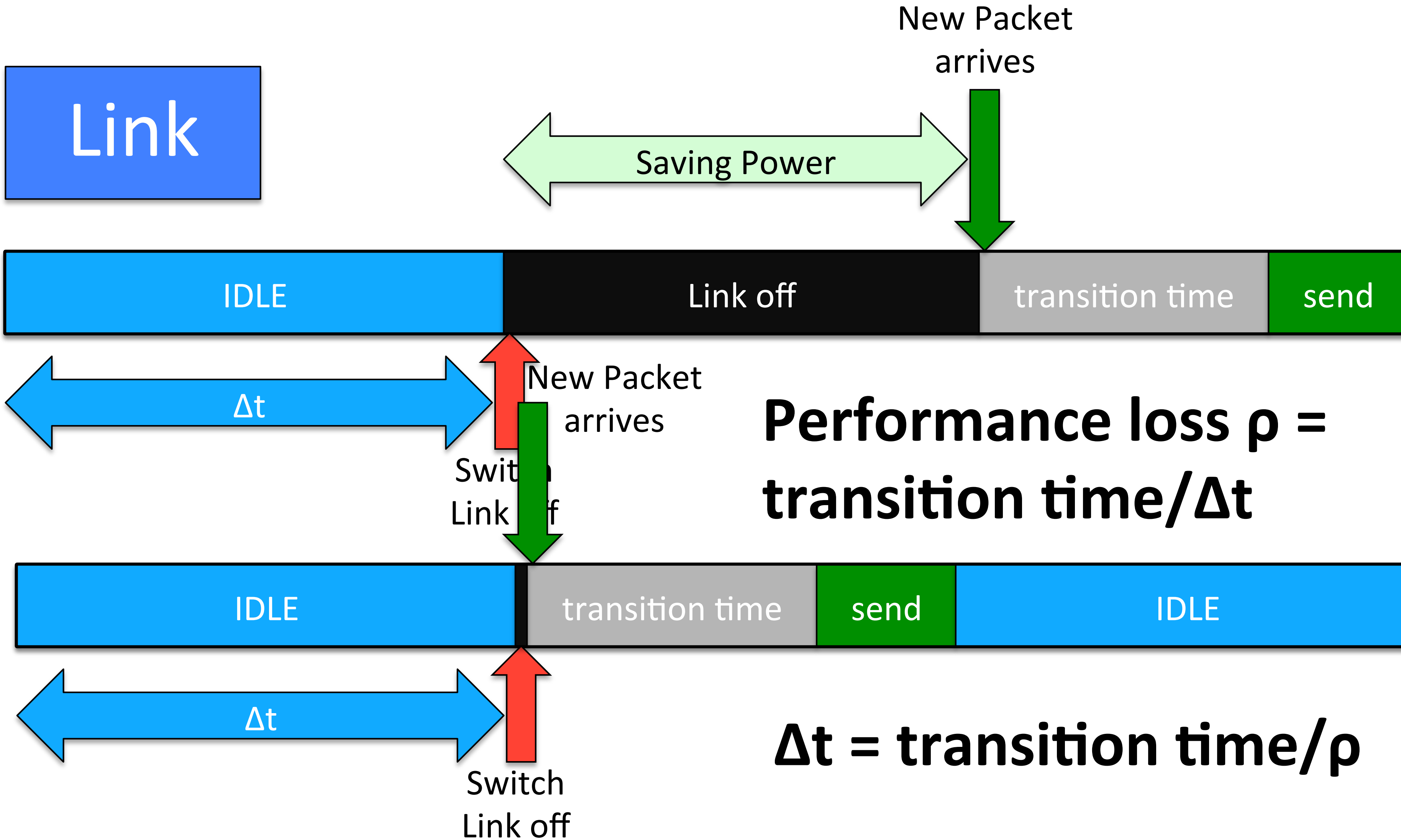
- Setup: OMNeT++-based simulator (SAURON)
  - 3D Torus topology
  - 512 nodes
  - XYZ-dimension-order routing



First naïve strategy

- First strategy limited to two different power states
- Saving power in interconnection networks by reducing bandwidth always correlates with performance loss
- Important parameters: transition time t_{trans} and max. performance loss ρ
- Links are switched off after idling for a certain time Δt
- Links are turned on again when a new packet arrives
  - A. Venkateshet al. “A case for application-oblivious energy- efficient mpi runtime,” *ISC*, 2015

First naïve strategy

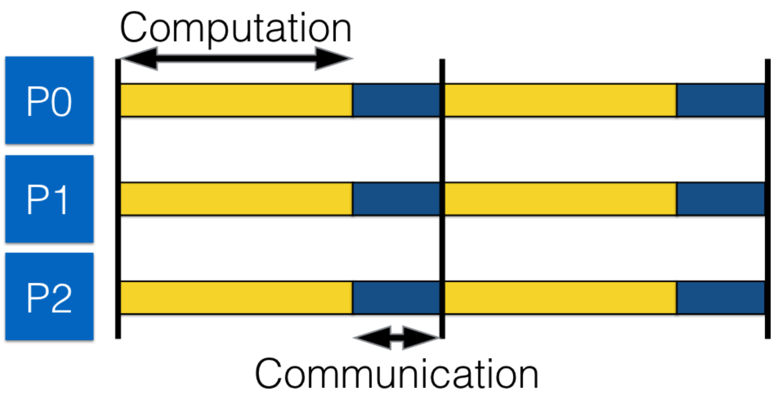


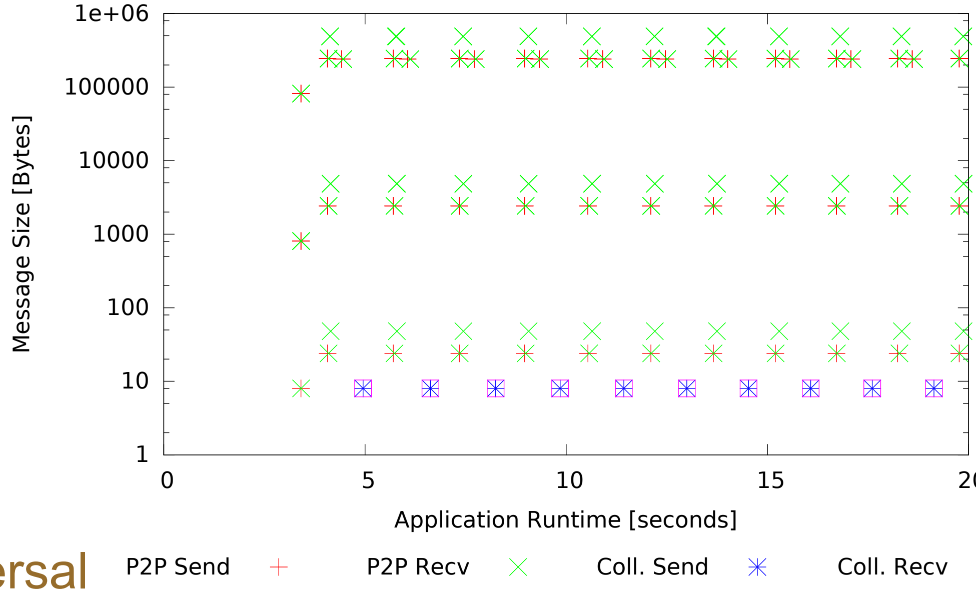
Performance loss ρ = transition time/Δt

Δt = transition time/ρ

Workloads – Selection (1)

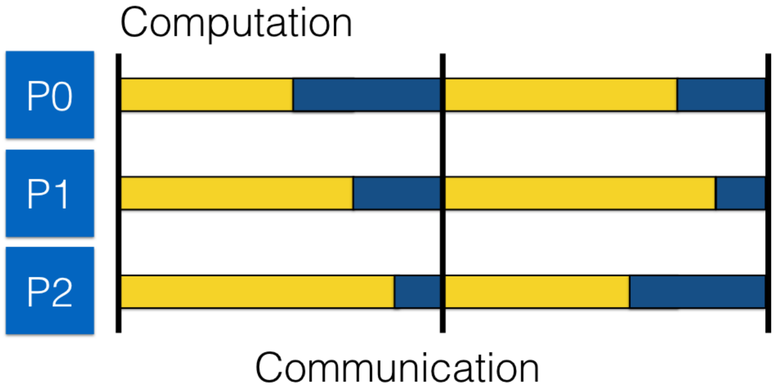
- LULESH
  - Very regular communication pattern
  - hydrodynamic simulations (stencil code)
  - One of DoE's proxy application for exascale computing
  - size = 100 (≅10^6 elements/node), iterations = 50
- NAMD
  - Iterative
  - molecular-dynamic application
  - STMV molecule (~10^6 atoms)
- Graph500
  - irregular
  - a breadth- first search graph traversal
  - scale factor = 20, edge factor = 16

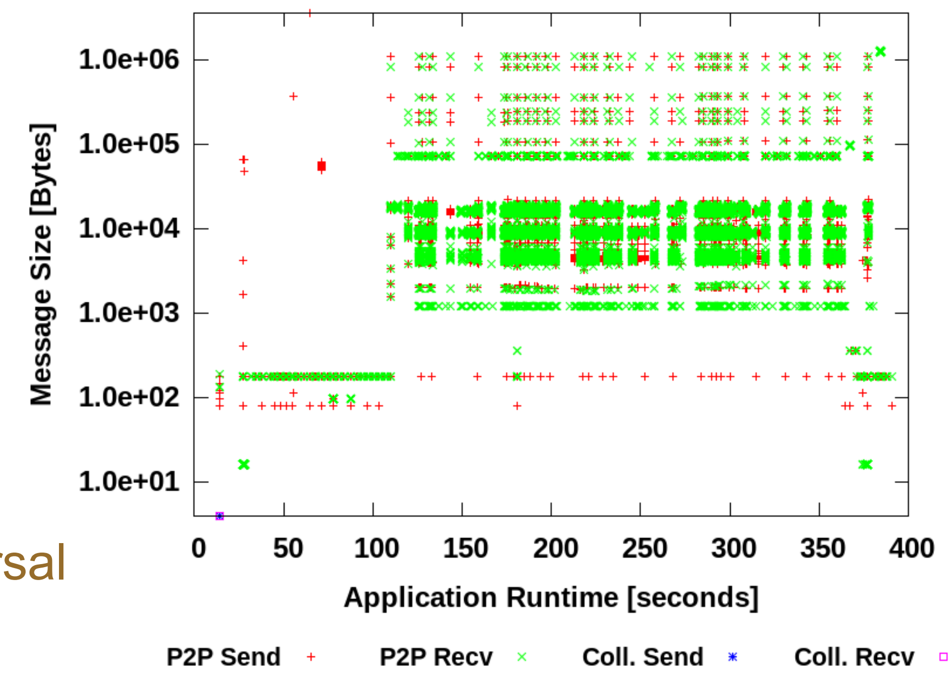




Workloads – Selection (2)

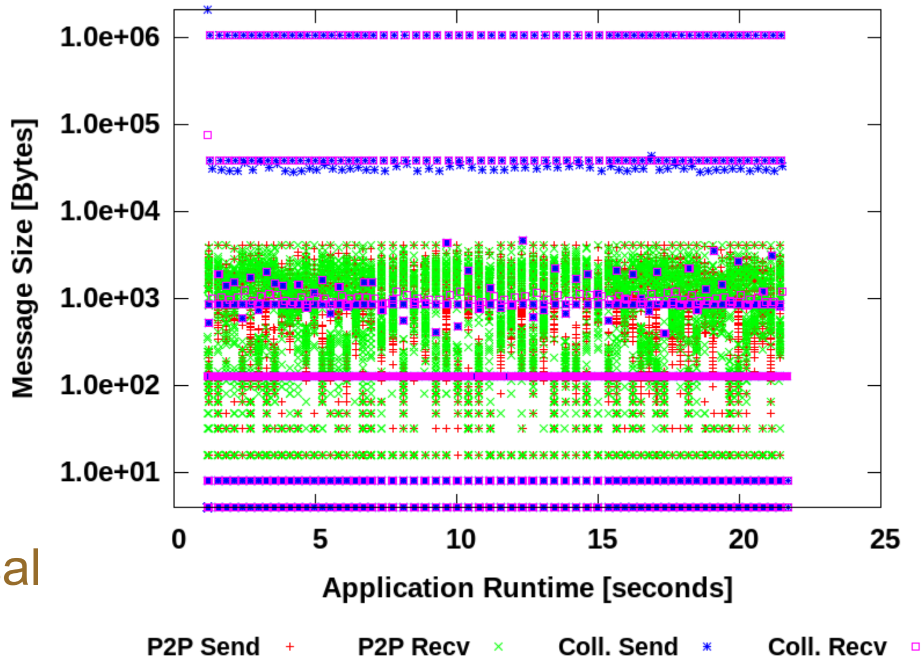
- LULESH
  - Very regular communication pattern
  - hydrodynamic simulations (stencil code)
  - One of DoE's proxy application for exascale computing
  - size = 100 (≅10^6 elements/node), iterations = 50
- NAMD
  - Iterative
  - molecular-dynamic application
  - STMV molecule (~10^6 atoms)
- Graph500
  - irregular
  - a breadth- first search graph traversal
  - scale factor = 20, edge factor = 16





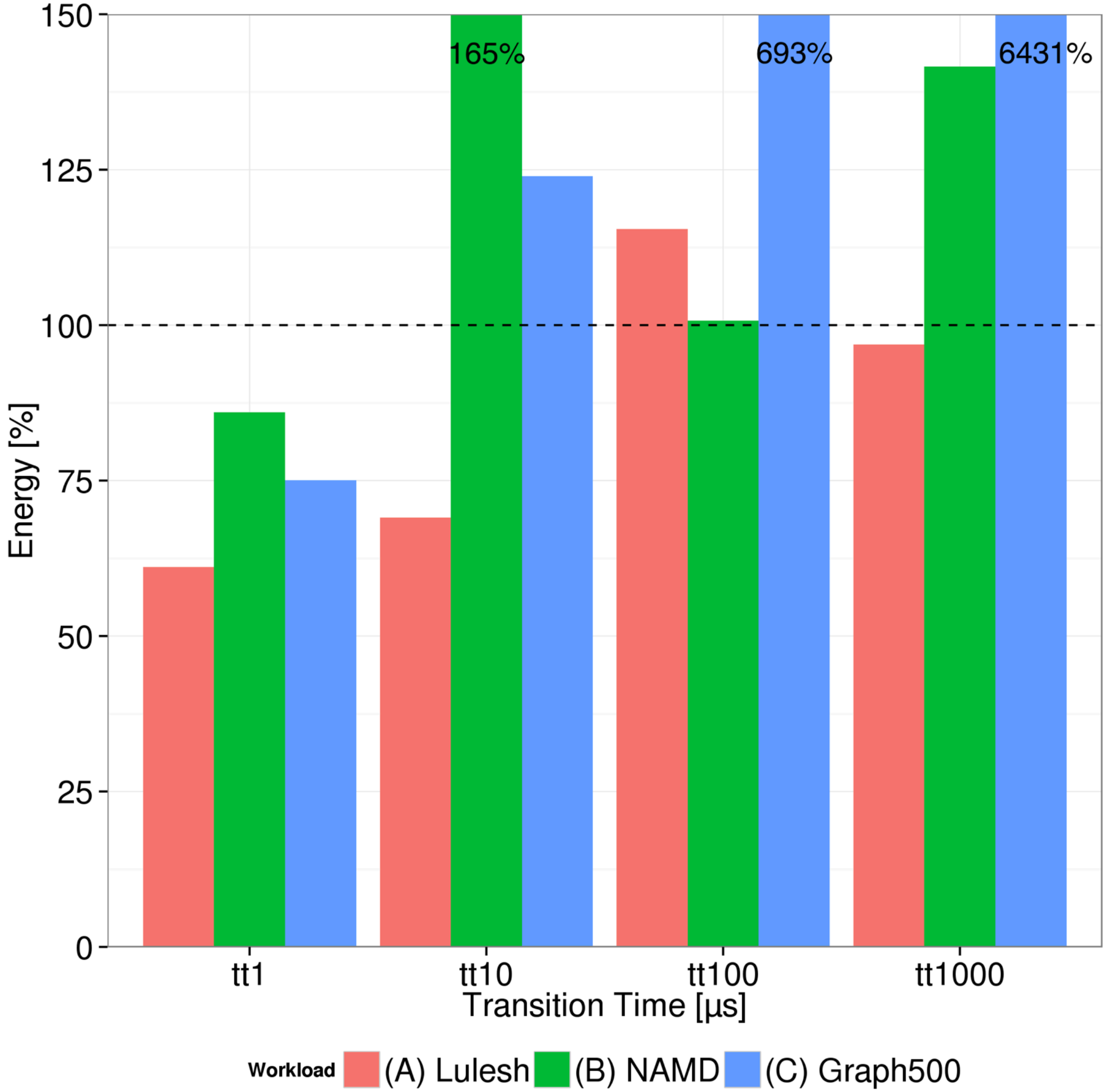
Workloads – Selection (3)

- LULESH
  - Very regular communication pattern
  - hydrodynamic simulations (stencil code)
  - One of DoE's proxy application for exascale computing
  - size = 100 (≅10^6 elements/node), iterations = 50
- NAMD
  - Iterative
  - molecular-dynamic application
  - STMV molecule (~10^6 atoms)
- Graph500
  - irregular
  - a breadth- first search graph traversal
  - scale factor = 20, edge factor = 16



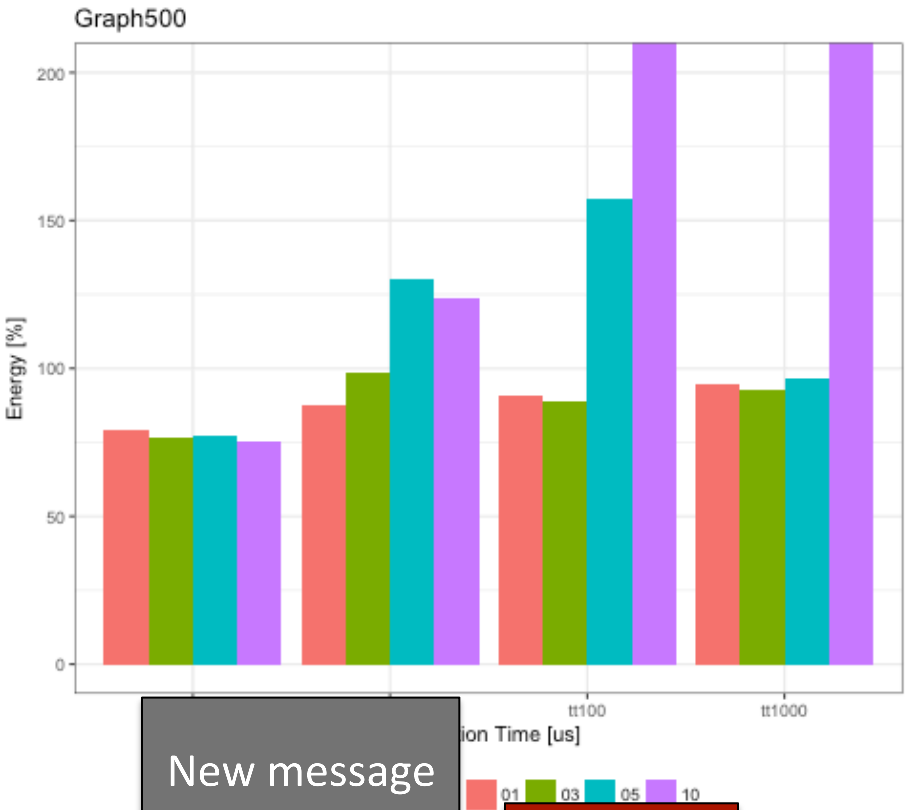
Results – Transition Time

- Max. perf. Loss: 10%
- Transition time has huge impact on power saving potential
- Longer transition times cause significant performance losses

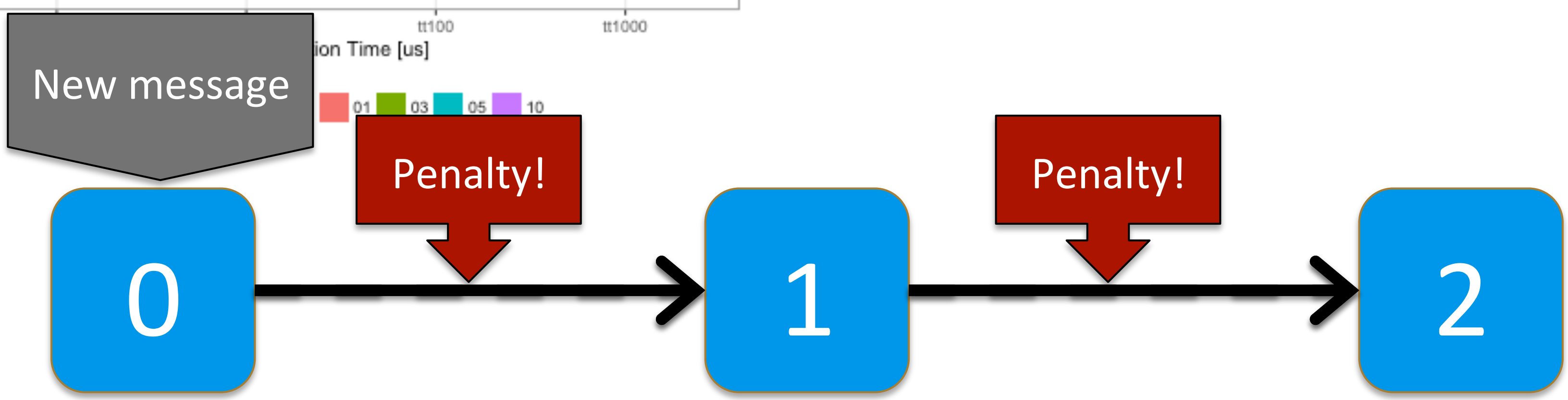


RUPRECHT-KARLS-
UNIVERSITÄT
HEIDELBERG

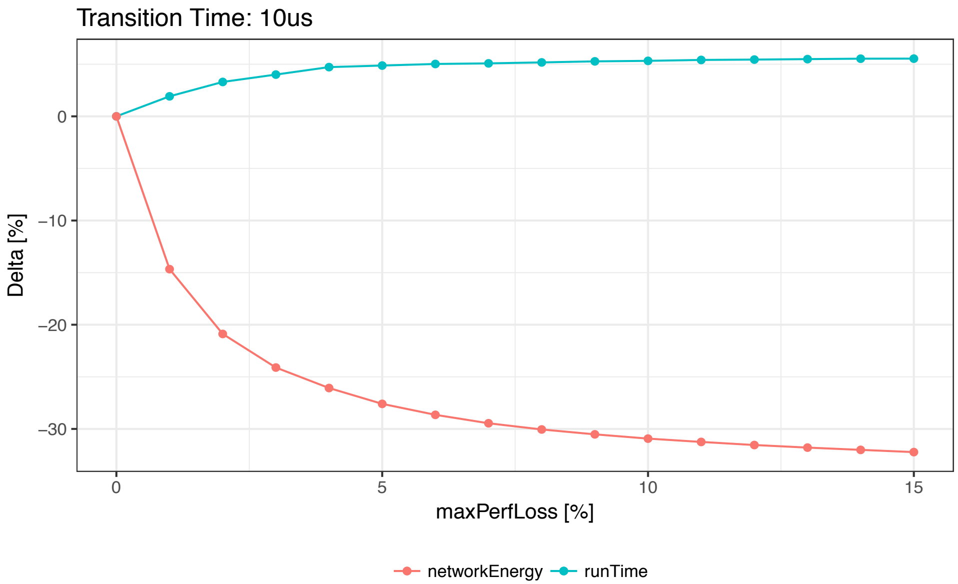
Insights - problems

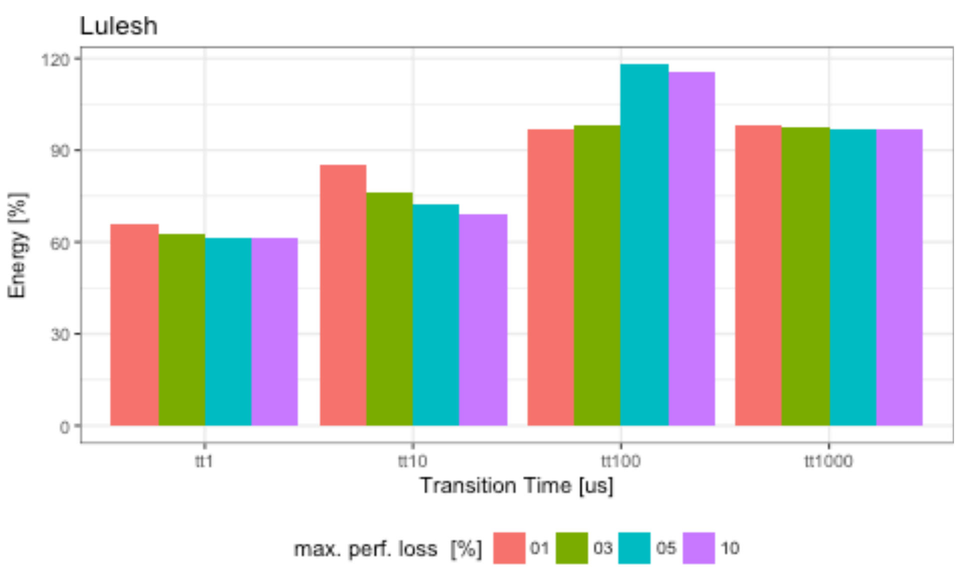


- Max. performance loss only from node view
- Penalties can accumulate in every node along a path



Results- LULESH

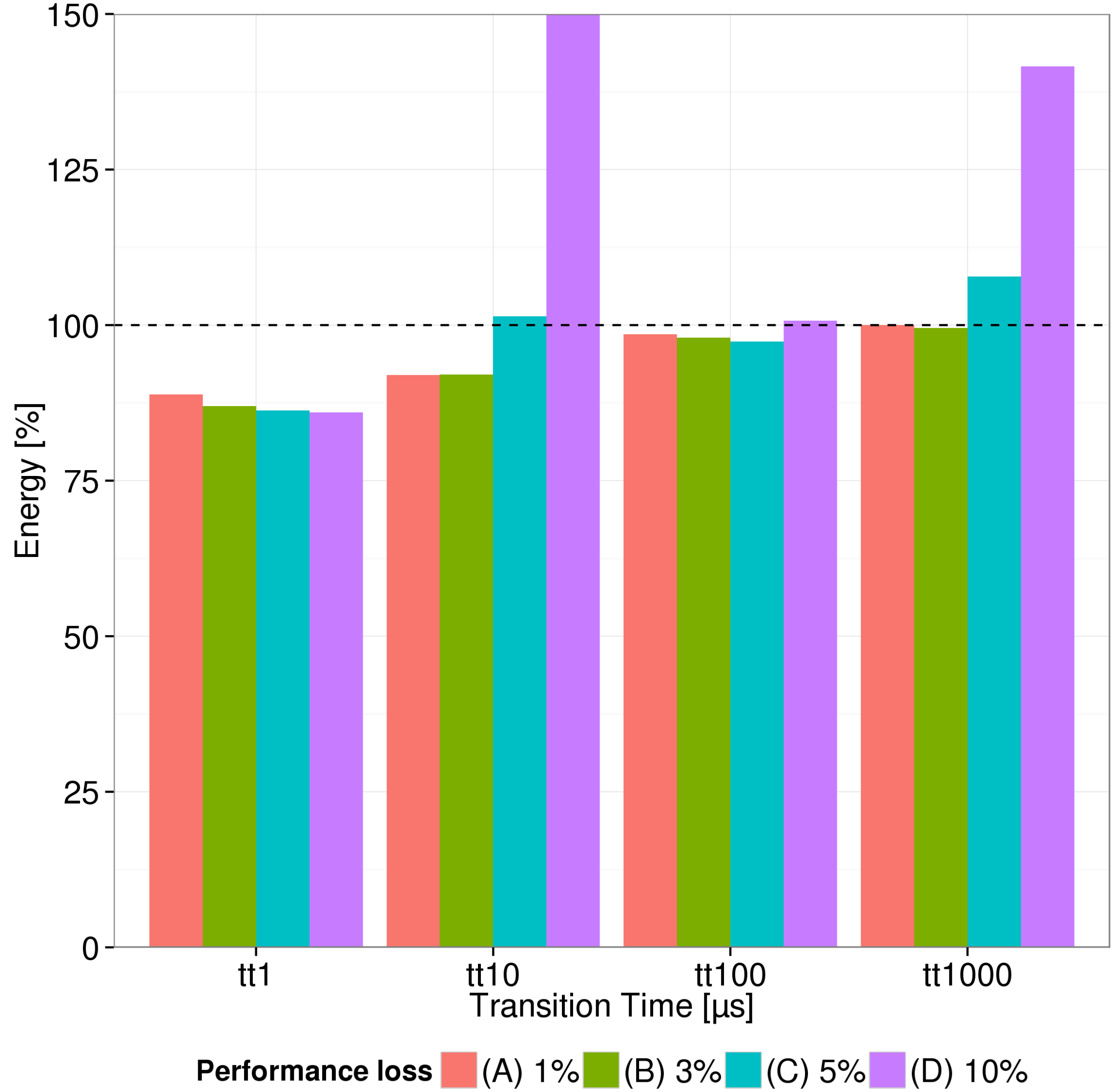




- Promising amount of energy was saved
  - At least for most configurations
- Most performance losses remain below the maximum
- Transition time is a crucial parameter for saving energy in the network

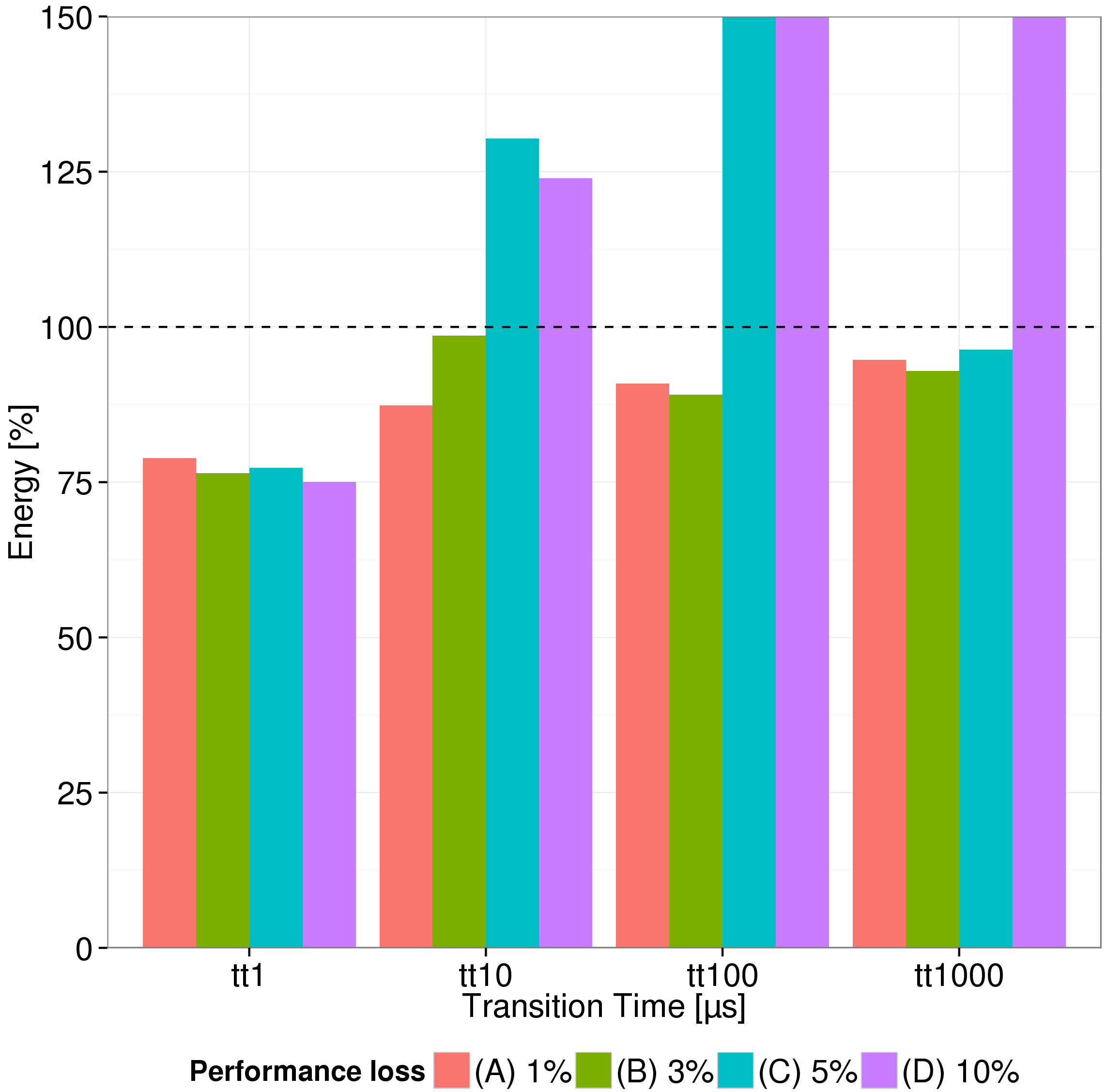
Results – NAMDstmv

- Less potential than LULESH
- Still energy savings in most configurations
- Worst results for short Δt



Results – Graph500

- Worst workload for energy saving purposes
- Although short transition times enable power saving potential



Summary

- Even simple On/OFF Policy enables much potential for energy/power saving
- Transition time is a crucial factor in this context
  - In future design decisions transition time should receive more attention
- Energy saving potential depends highly on communication pattern
  - The more regular the communication pattern, the more potential
  - Observations suggest that there will be no “perfect” policy
  - Instead different policies for different communication pattern

Future Work

- More Topologies, more Workloads
- Better Policy:
  - Instead of switching links off, moving to lowest power state that is able to send data => less power saving, better perfomance
  - Introducing an “awake message”, which wakes up all links along a path while message needs to wait for the first link to reconfigure
  - Measuring utilization in links. If a link is highly utilized by small messages, it is switched on again
- Using congestion management or alternative routing strategies in order to minimize performance losses
- Moving power saving from Port to System view in order to keep the big picture in eye

Credits

Discussions: Benjamin Klenk, Alexander Matz, (Heidelberg University), Francisco Andujar (Universitat Politècnica de València) , Pedro J. Garcia Jesus Escudero, Pedro Yebenes (Universidad de Castilla-La Mancha)

Current main interactions

UCLM
UNIVERSIDAD DE CASTILLA~LA MANCHA

Thank you!

Questions?