Communication Forecasting for Large-Scale Applications

Nikela Papadopoulou, Georgios Goumas, Nectarios Koziris

Computing Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens

Motivation

- The greatest challenge in HPC is to attain petaflop performance on a petaflop machine - and exaflop will be harder!
- Users tend to make overly optimistic assumptions about their applications' scalability - execution results refute their expectations
- A key to attain maximum parallel performance is **predictive modeling** - we need to know what to expect!
- Effective **performance models** for large-scale parallel applications can be a valuable tool for **decision-making** at many levels:
  - allocation and utilization of resources by users
  - job scheduling on large-scale systems
  - code optimizations (e.g. hybrid MPI/OpenMP, message compression) for performance tuning and performance portability
  - performance auto-tuning at runtime

Challenges

- HPC applications have computation and communication phases.
- Communication on large scale is affected by:
  - the data volume
  - the communication pattern
  - the programming model and communication primitives
  - the node architecture
  - the network architecture, protocols and topology
  - the process mapping on the allocation
  - ... who knows what else!
- Computation forecasting seems trivial compared to communication!

Approach

- Topology-agnostic modeling (trade accuracy for generality)
- Application-related model variables (easy to extract)
- Architecture-related model coefficients (hide complexity)
- Prediction of the per-phase communication time
- Supervised learning (coefficient training with benchmark results)

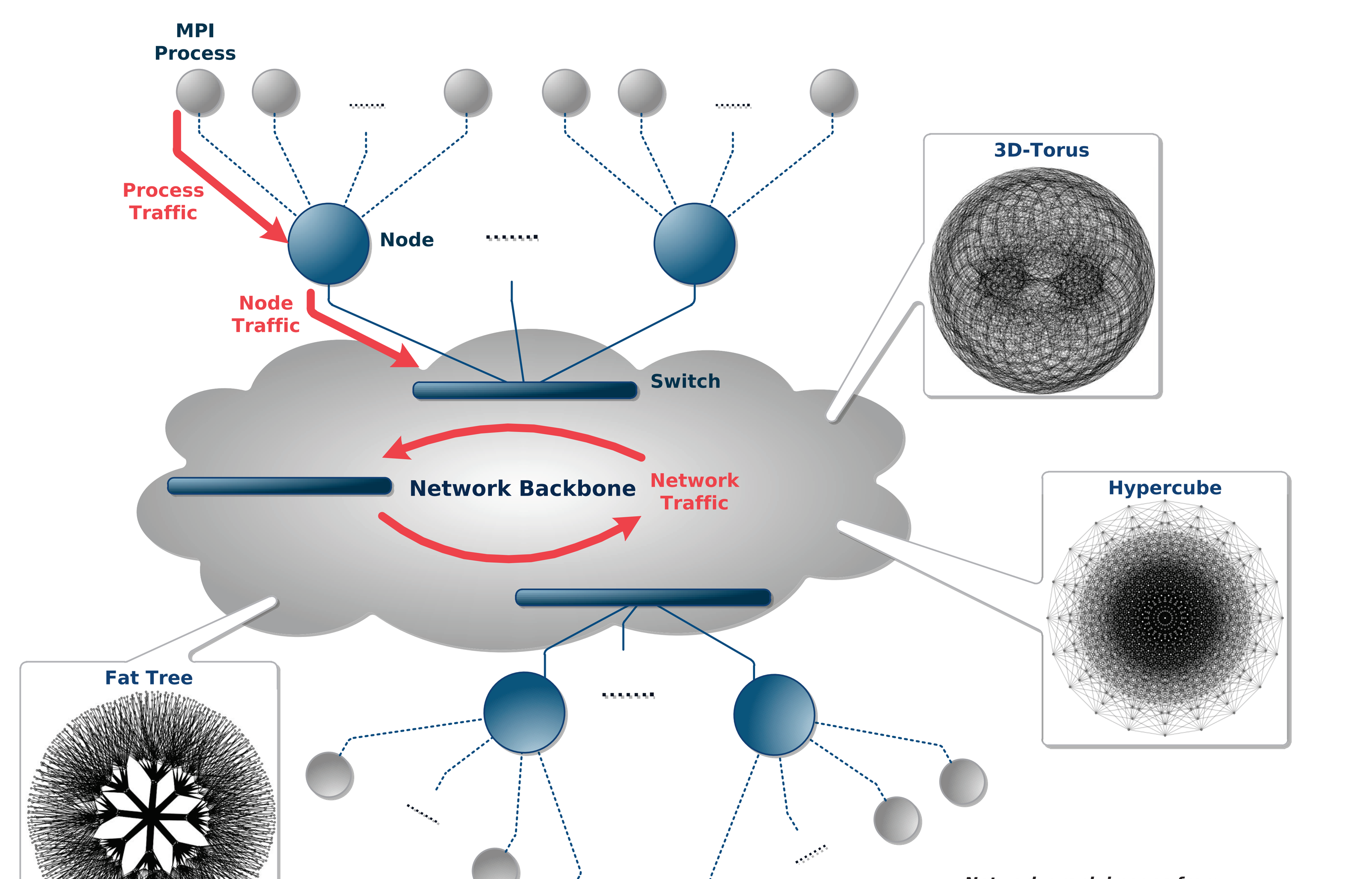
Methodology

1. **Benchmarking** of the system
2. **Variable selection** with statistical analysis
3. **Model building** with forward stepwise regression
4. **Multiple variable regression** to compute model coefficients
5. Refinement of the model

Execution Environment

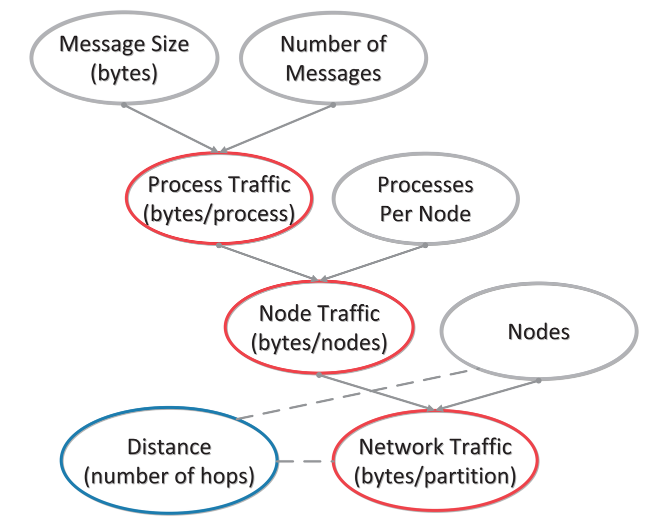
We experimented on **Vilje supercomputer** at NTNU (#82 at Top500), an SGI system of 1404 Intel Xeon E5-2670 dual eight-core nodes interconnected with Infiniband FDR on an enhanced hypercube.

Communication Data Flows



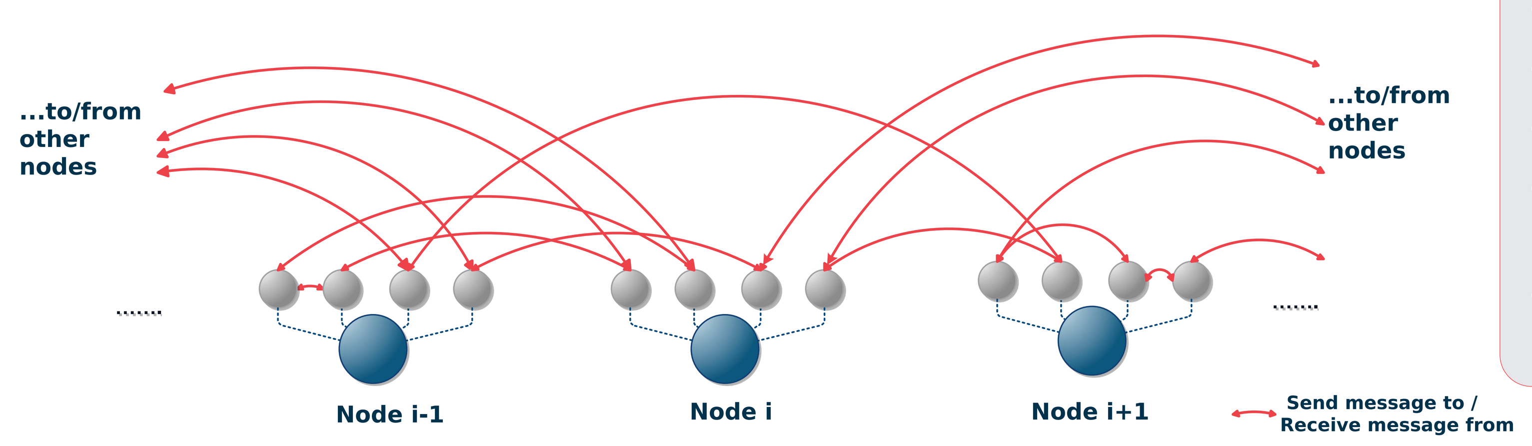
Network graph images from:
1. http://htor.inf.ethz.ch/research/topologies/ (3D Torus, Fat Tree)
2. http://en.wikipedia.org/wiki/Hypercube (Hypercube)

Performance Metrics



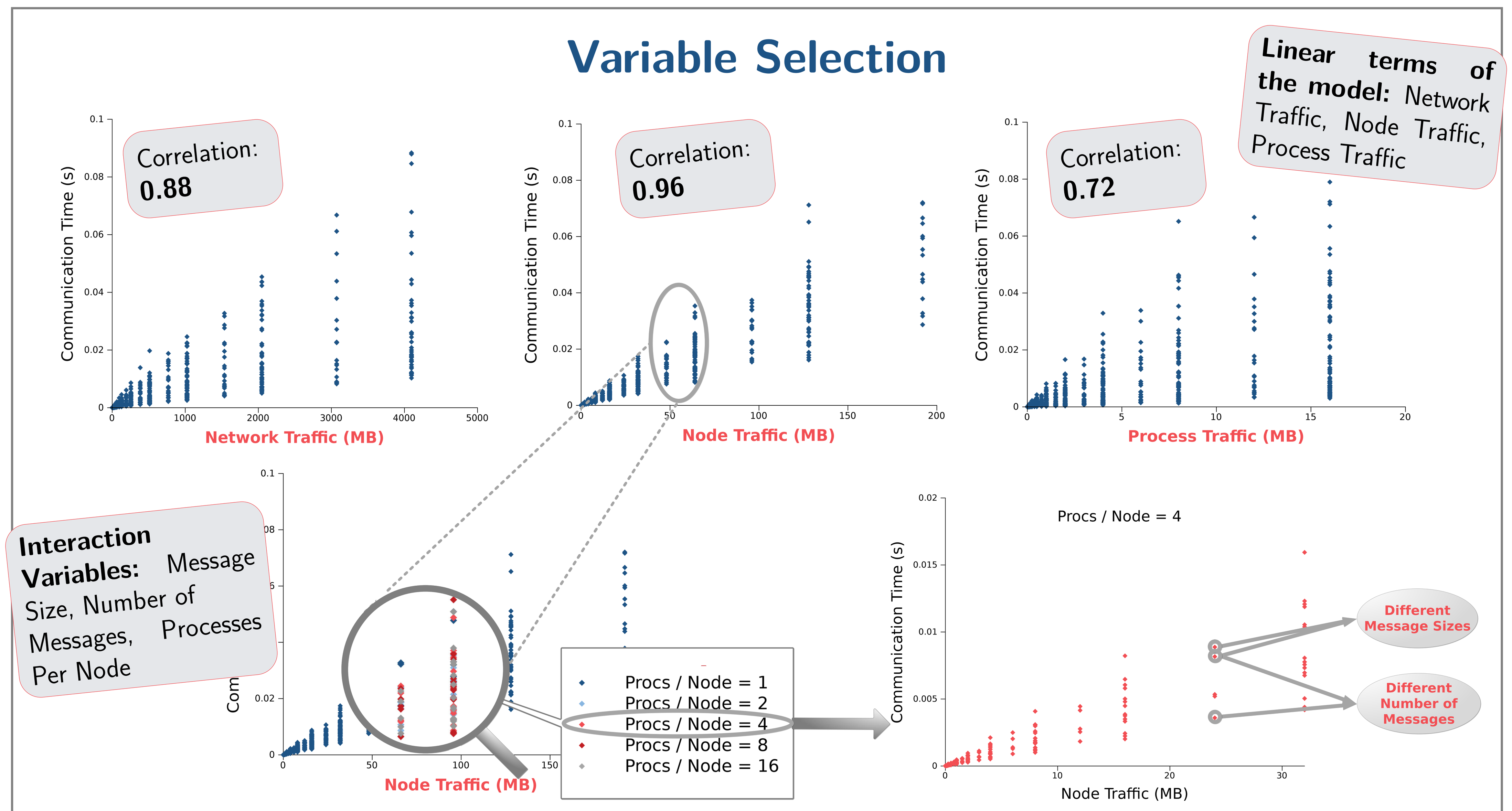
- Traffic - **parallelism** or **congestion**
- Message size - **protocol changes**
- Distance - **contention** or **parallelism**
- Processes per node - **overlapping capacity** or **contention**

Benchmarking

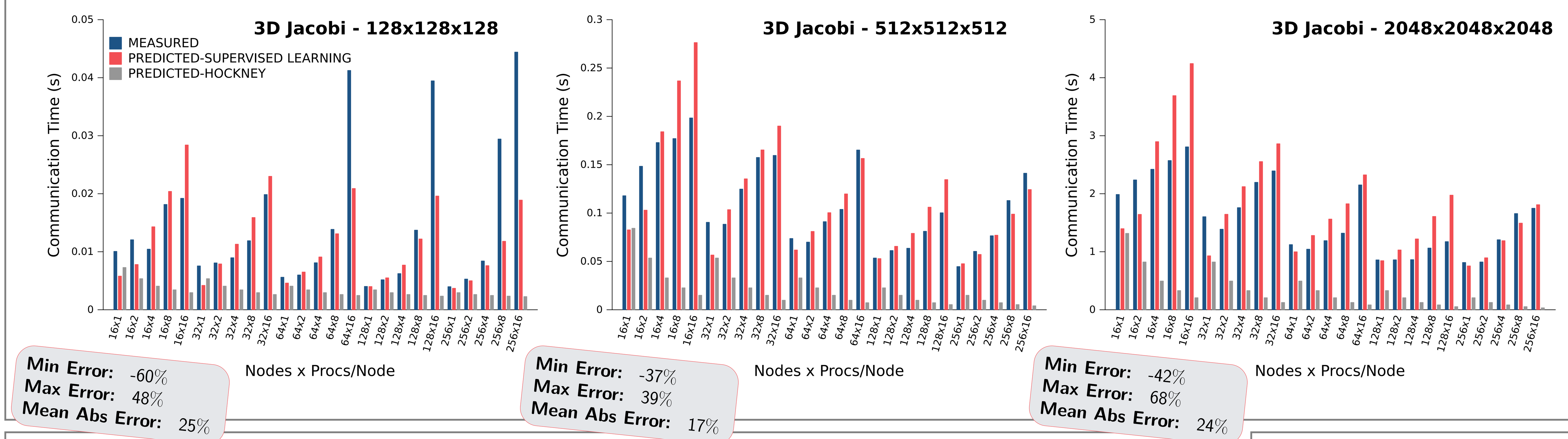


- Non-blocking ping-pongs with MPI
- Multiple random pairs of processes
- Multiple messages per process
- Various configurations of nodes and processes per node
- Random mixture of internode/intranode communication

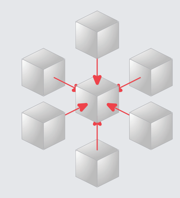
Variable Selection



Evaluation



For the evaluation of our supervised-learning performance model, we executed a parallel 3D Jacobi, an iterative PDE solver. 3D Jacobi has a 5-point stencil communication pattern with halo exchanges of the 2D-faces.



Computation and communication phases are discrete and communication is orchestrated with MPI point-to-point non-blocking communication primitives. We compared the actual communication times against the communication times predicted with our supervised-learning model and with the Hockney model:

$$t_{comm} = Latency + MessageSize/Bandwidth$$

Future Work

- Improve model accuracy, by studying and modeling more complex features of communication
- Experiment on different systems and network topologies
- Communication forecasting for applications with more intricate communication patterns
- Extend methodology for collective communication
- Automate the performance modeling process and construct a generic, portable tool for performance prediction

Acknowledgements

This research was partly funded by project I-PARTS (code 2504) of Action ARISTEIA, co-financed by the European Union (European Social Fund) and Hellenic national funds through the Operational Program Education and Lifelong Learning' (NSRF 2007-2013). We would like to thank NTNU, the Norwegian University of Science and Technology at Trondheim, Norway, for granting us access to Vilje Supercomputer.

LaTeX TikZposter