Towards a Performance-as-a-Service Cloud
D. B. Bartolini,* F. Sironi,* M. Maggio,+ G. C. Durelli,* D. Sciuto,* M. D. Santambrogio*
*Politecnico di Milano, +Lund University

Background and Motivation
IaaS clouds have a resource-based interface, but users care about performance: there is a fundamental mismatch.

- Users have to fine-tune allocations per-VM
- Providers need to maintain consistent performance despite resource sharing
- The fine-grained and fast resource trading scenario that is emerging will magnify these issues

Contribution
Build a Performance-as-a-Service (PeaaS) model where users state service-level objectives (SLOs) for performance-sensitive VMs

1) dispense with per-VM allocation fine tuning
2) enable providers to optimize infrastructure utilization, supporting batch best-effort workloads
3) take advantage of fine-grained, high-frequency resource trading

Initial Implementation and Results on CMP Allocation
We use a two-level control schema, with application-level controllers and node-level resource brokers

We initially focus on chip multiprocessor (CMP) allocation to co-located compute-bound workloads; we evaluate a prototype on managing performance-sensitive (1) swaptions and (2) x264 from the PARSEC 2.1 suite, co-located on a six-core CMP with a batch best-effort workload (SPECjbb2005), that we use to maximize node utilization

Conclusion
Our prototype dynamically manages CMP allocation so as to meet SLOs and maximize node utilization

Our lightweight controllers run at high-frequency (sub-second period) and can react to fast variations that may cause long tail latency distributions

Work in Progress / Open Issues
- We are extending our prototype to support more resources (i.e., I/O bandwidth, cache)
- We will evaluate our extended system with representative multi-tier cloud workloads
- We are studying broader issues, such as fair pricing in the PeaaS model