D3

Invading instead of Wasting HPC resources

Michael Gerndt
Technische Universität München

10th VI-HPS Aniversary, Frankfurt

D3

Invasive Resource Management

Dynamic

- Applications
- Capabilities
- Resources

Static Resource Management

Invasive Resource Management

D3

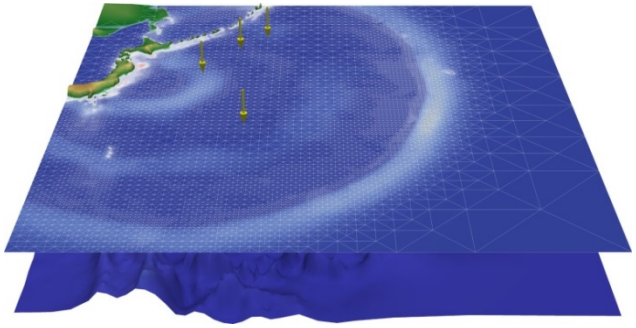
Scenarios:

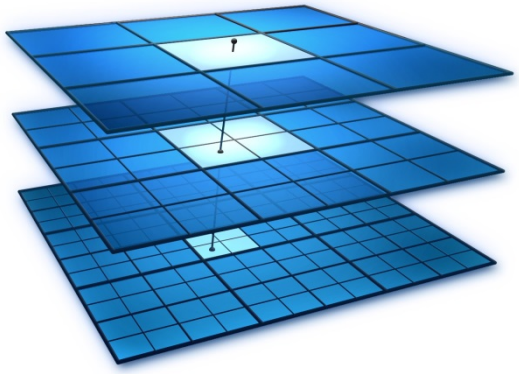
- Phase-based applications
- Adaptive applications
- Jobs with different input data set
- Analytics applications with data bursts
- Coupled applications with dynamically varying resource requirements
- IO intensive phases require network and IO bandwidth
- Node failures
- Urgent computing
- Power stability
- Increased scheduling opportunities
- …

InvasIC: D3 Subproject

D3

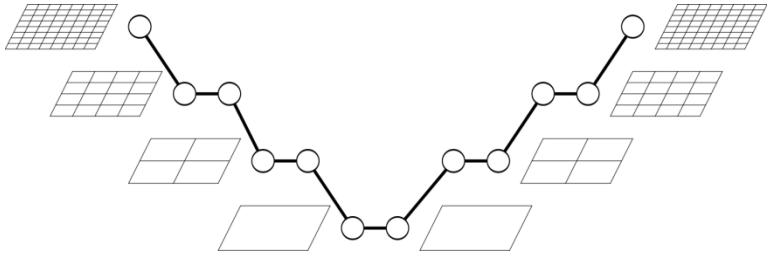
Invasive HPC applications

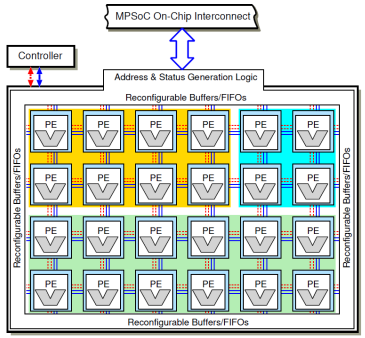




Invasive resource management for HPC

Invasive X10 applications on MPSoCs





Resource Adaptivity for MPI Applications

Implementation

- **MPI extensions**
  - Extend the API with adaptive operations
- **MPI Library**
  - Based on MPICH 3.2
- **Resource Manager**
  - Based on SLURM 15.08

MPI Extensions Overview

D3

Proposed 4 new operations as an extension to the MPI standard:

`MPI_Init_adapt`(…)
- Initializes the library in adaptive mode

`MPI_Probe_adapt`(…)
- Probes the resource manager for adaptations

`MPI_Comm_adapt_begin`(…)
- Marks the beginning of an adaptation window
- Provides inter communicator and new communicator

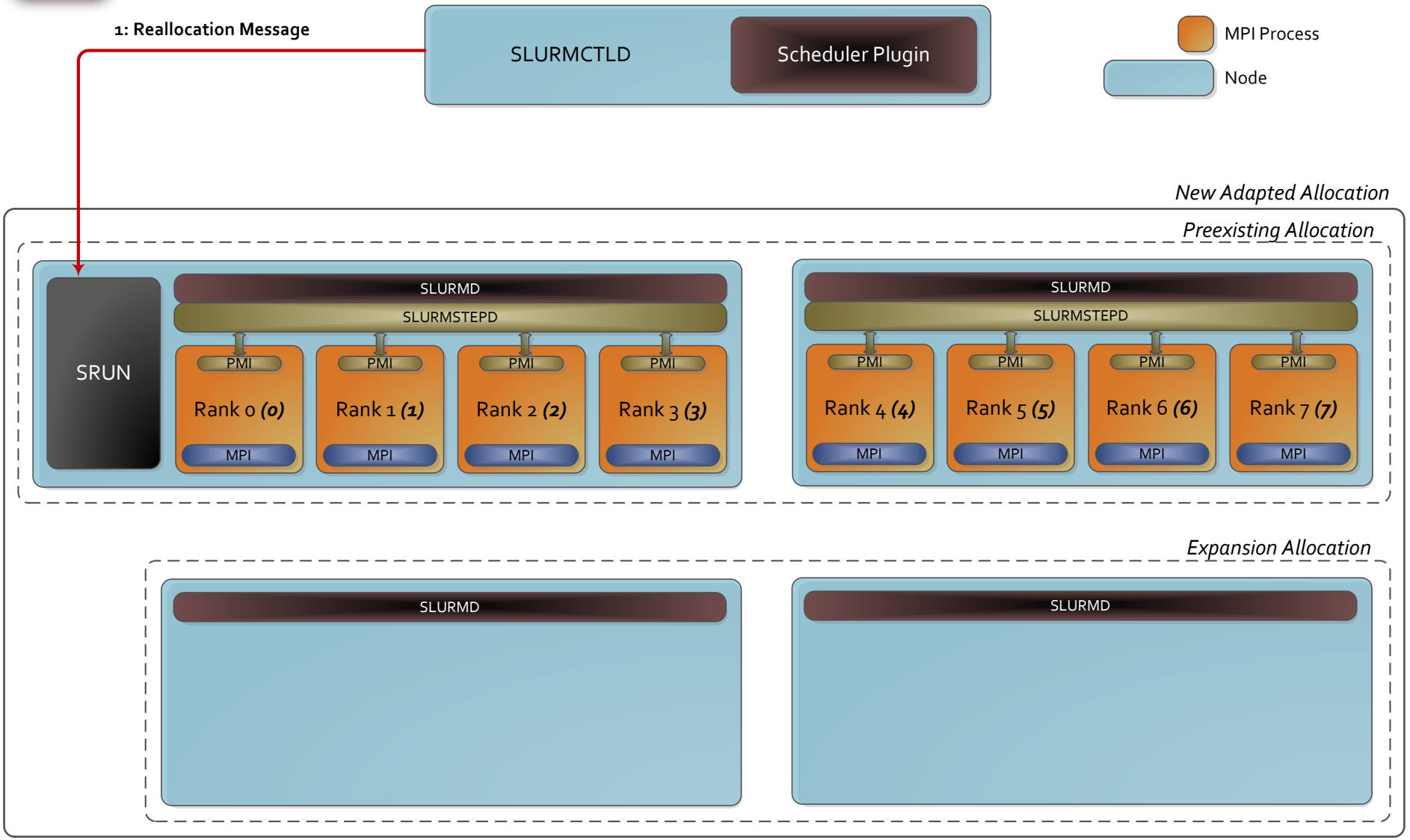
`MPI_Comm_adapt_commit`(…)
- Marks the end of an adaptation window
- Sets adapted `MPI_COMM_WORLD`

Code Structure

```
MPI_Init_adapt(…, &status);
for (…) {
    MPI_Probe_adapt(&adapt,…);
    if(adapt){
        MPI_Comm_adapt_begin(…);
        // redistribution code
        MPI_Comm_adapt_commit(…);
    }
    // compute and MPI code
}
```

Adaptation Step 1

D3



Adaptation Step 2

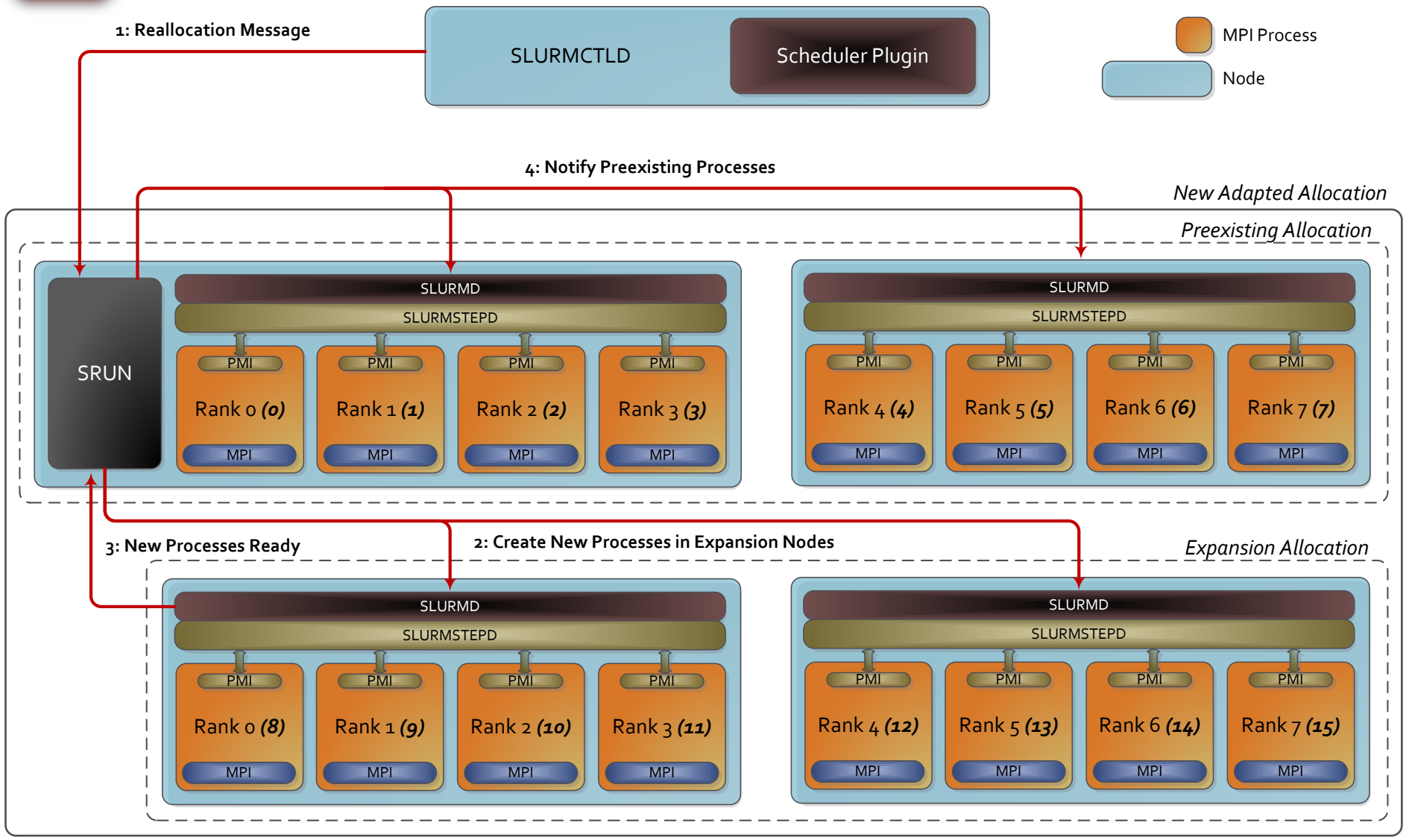
D3

Adaptation Step 3

D3

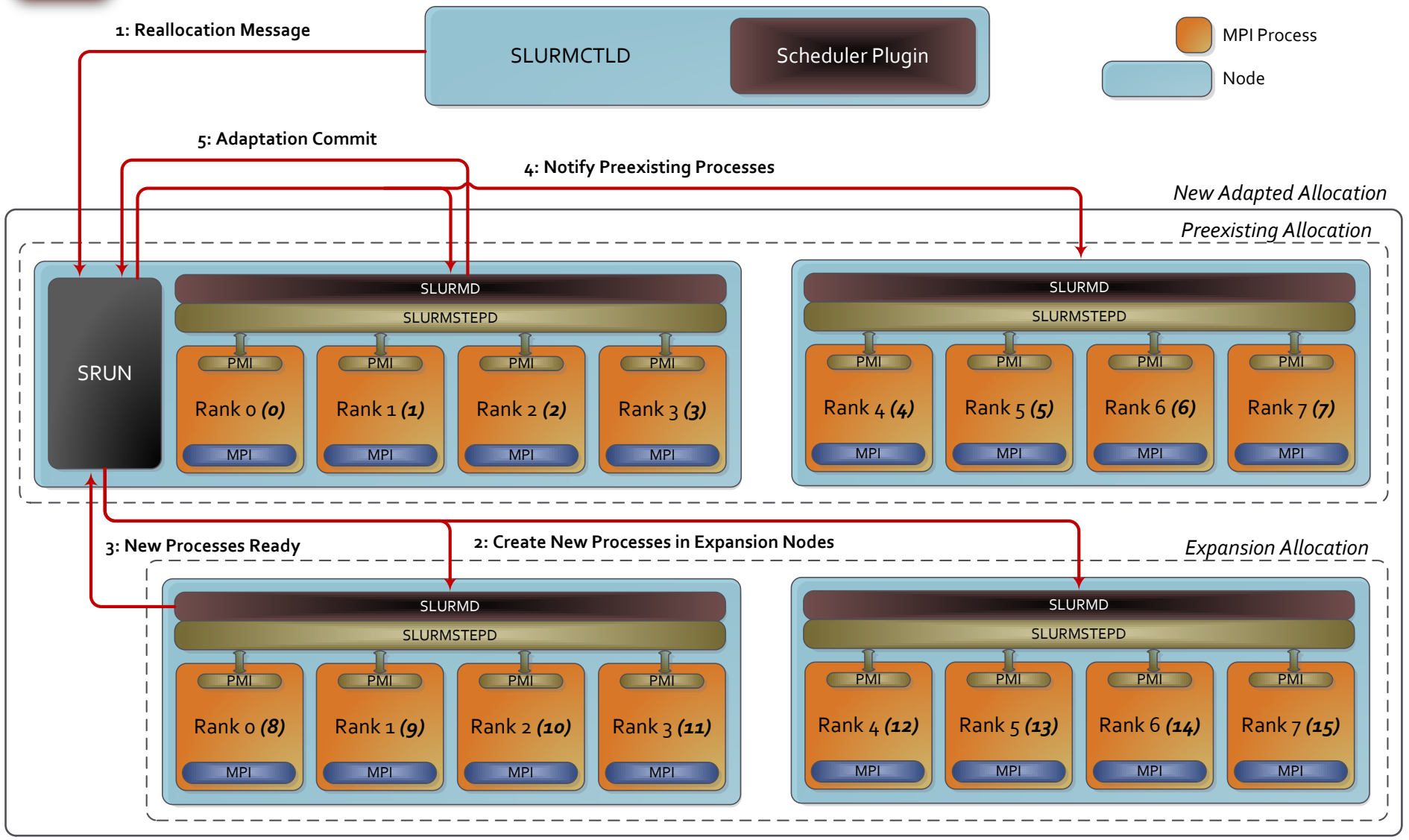
Adaptation Step 4

D3



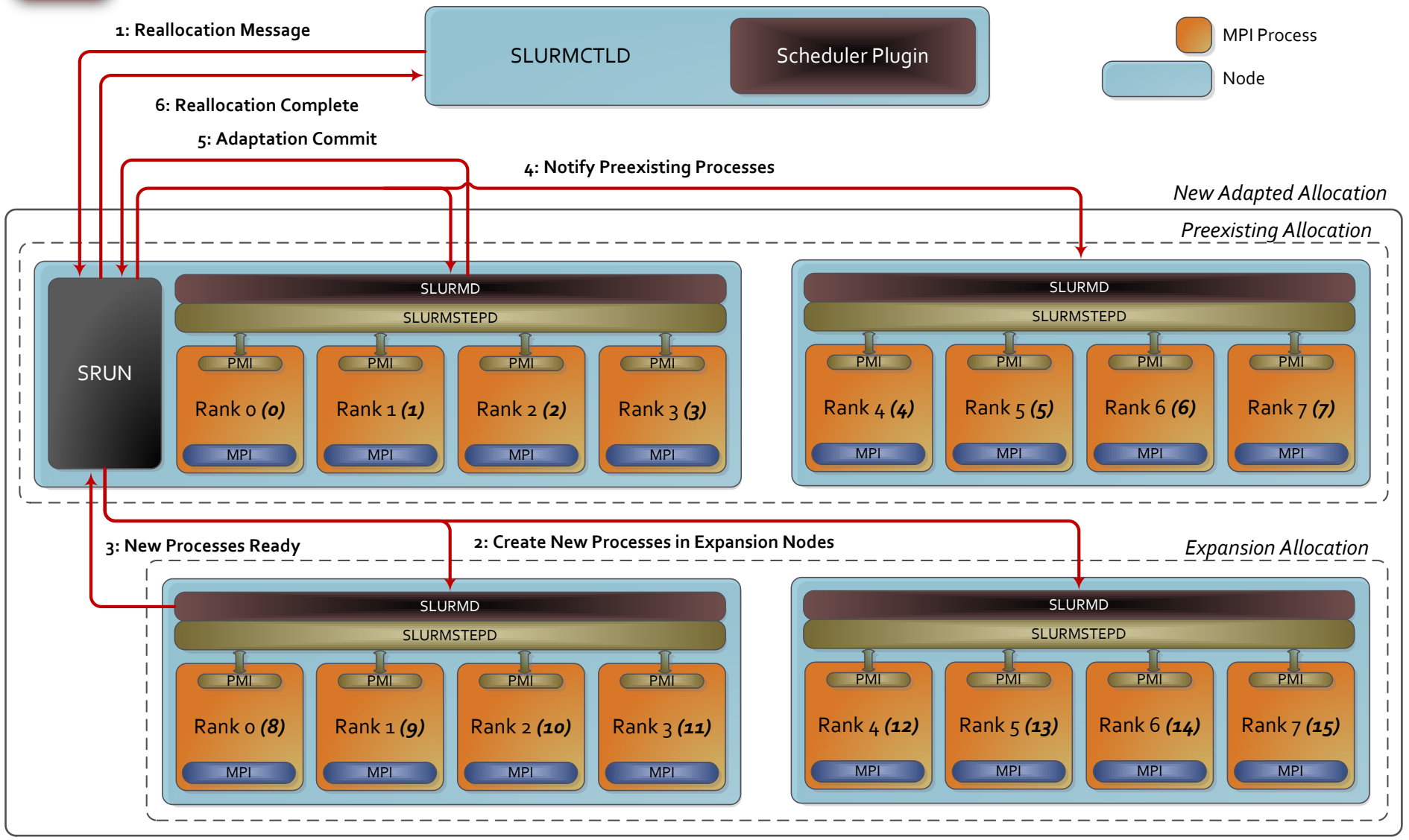
Adaptation Step 5

D3



Adaptation Step 6

D3



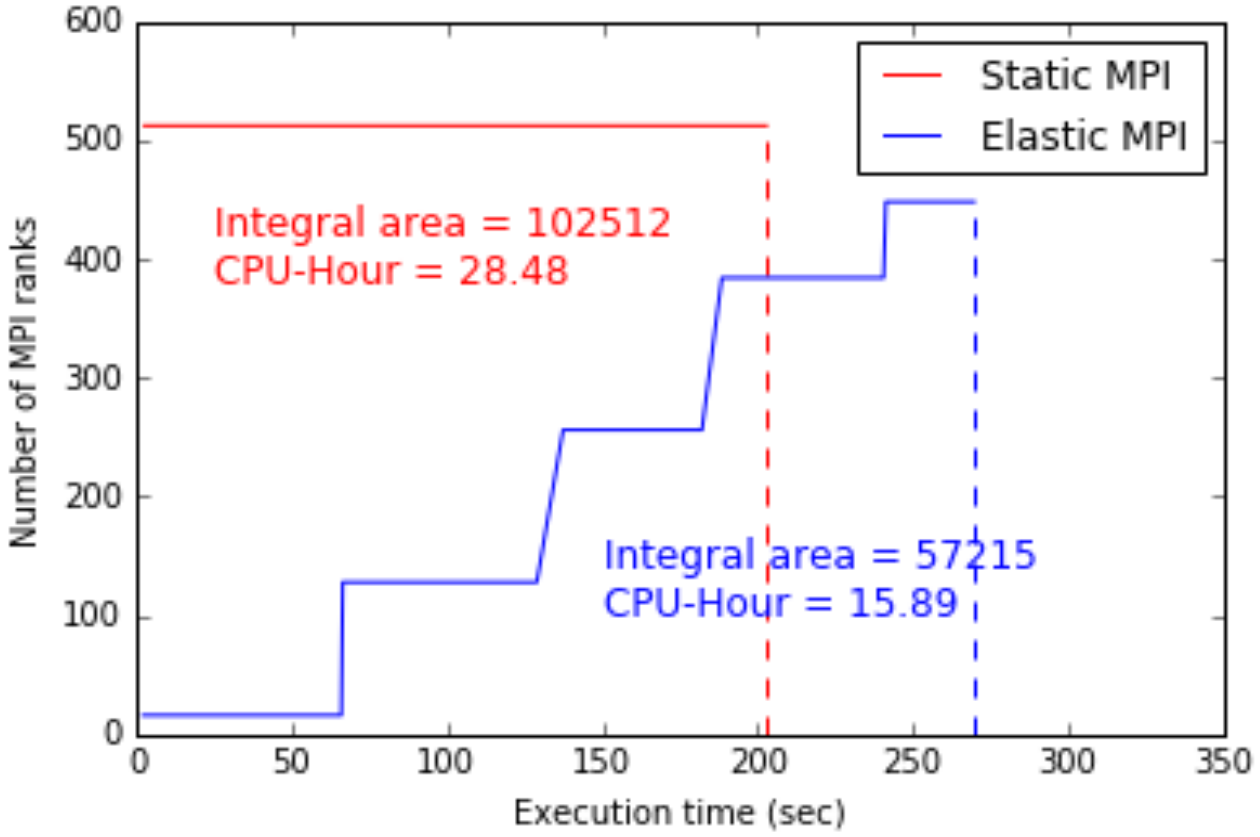
Experiment Platform

D3

SuperMUC @ LRZ

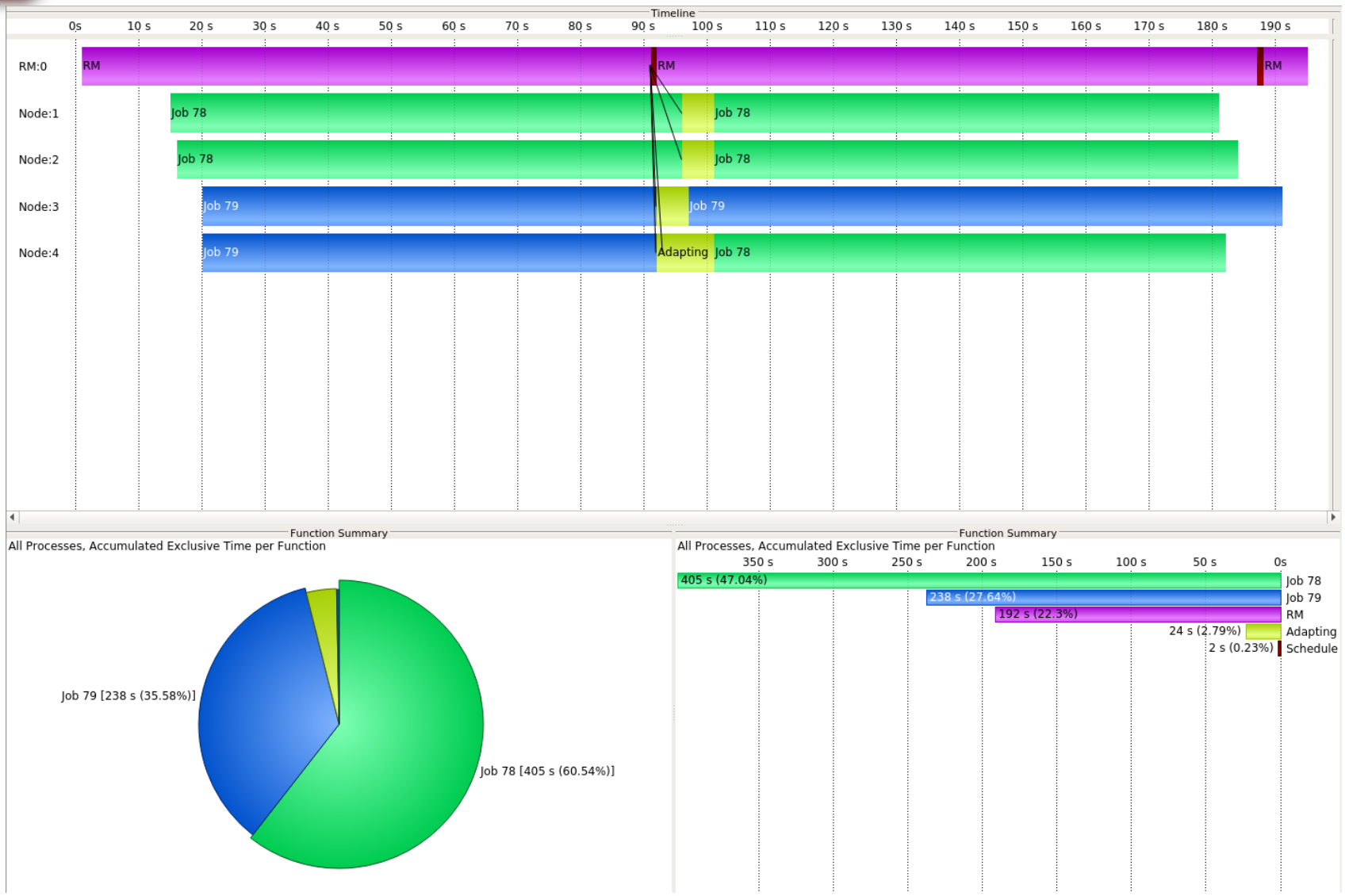
- Allocation of set of SuperMUC nodes via batch job
- Management of the nodes via separate SLURM instance
- Distribution of resource management into
  - SLURM Scheduler
    - Selection and scheduling of invasive jobs
    - Based on resource offer
  - Invasic Scheduler
    - Invasive resource management
- Submission of new invasive jobs through sbatch command

Tsunami simulations (Mo-Hellenbrand, Bungartz)



Adaptation Visualisation

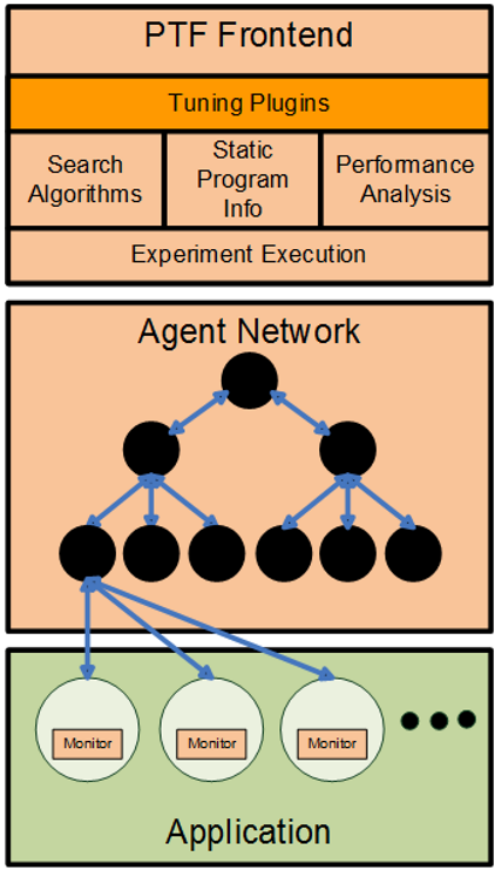
D3



Periscope Tuning Framework

VI-HPS, AutoTune, READEX

- Periscope Frontend
  - Controls the analysis and tuning process
  - Performs a sequence of experiments
    - While the application is executing
    - Based on application phases
    - Automatically starting/restarting the application if required
- Agent Network for scalability
  - Leave agents responsible for a subset of the MPI processes
  - Intermediate agents aggregate performance properties
- Online Access to the monitoring system
  - Configuration of measurements and tuning actions
  - Retrieval of performance data



Thank You