Reducing Speed for Energy Saving: Using RAPL Powercapping in HPC Systems.

Koudş HALITIM

Univ. Grenoble Alpes, Inria, Ctrl-A, CNRS, Grenoble INP, LIG, France

Supervisors :

Sophie Cerf, Eric Rutten, Raphael Bleuse, Bogdan Robu, Alexandre Van Kempen

Agenda :

1. Context :

   - HPC systems temporal behavior and energy efficiency
   - Targeted HPC system architecture

2. Approach and methodology :

   - RAPL powercapping on HPC
   - Feedback loop formulation

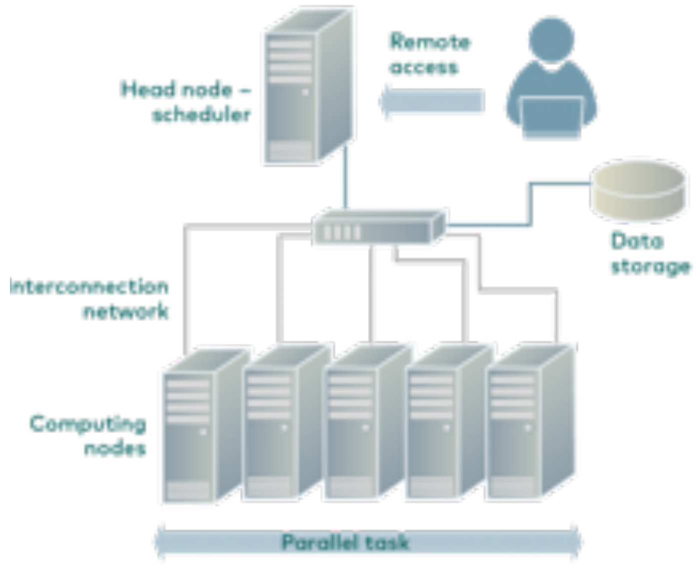
3. Modeling and control :

   - Cascaded Control: Addressing RAPL inaccuracies.
   - Evaluating the controller performance.

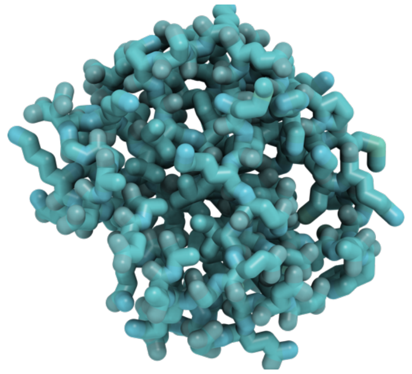
4- Takeaways

Context

HPC Systems [1]

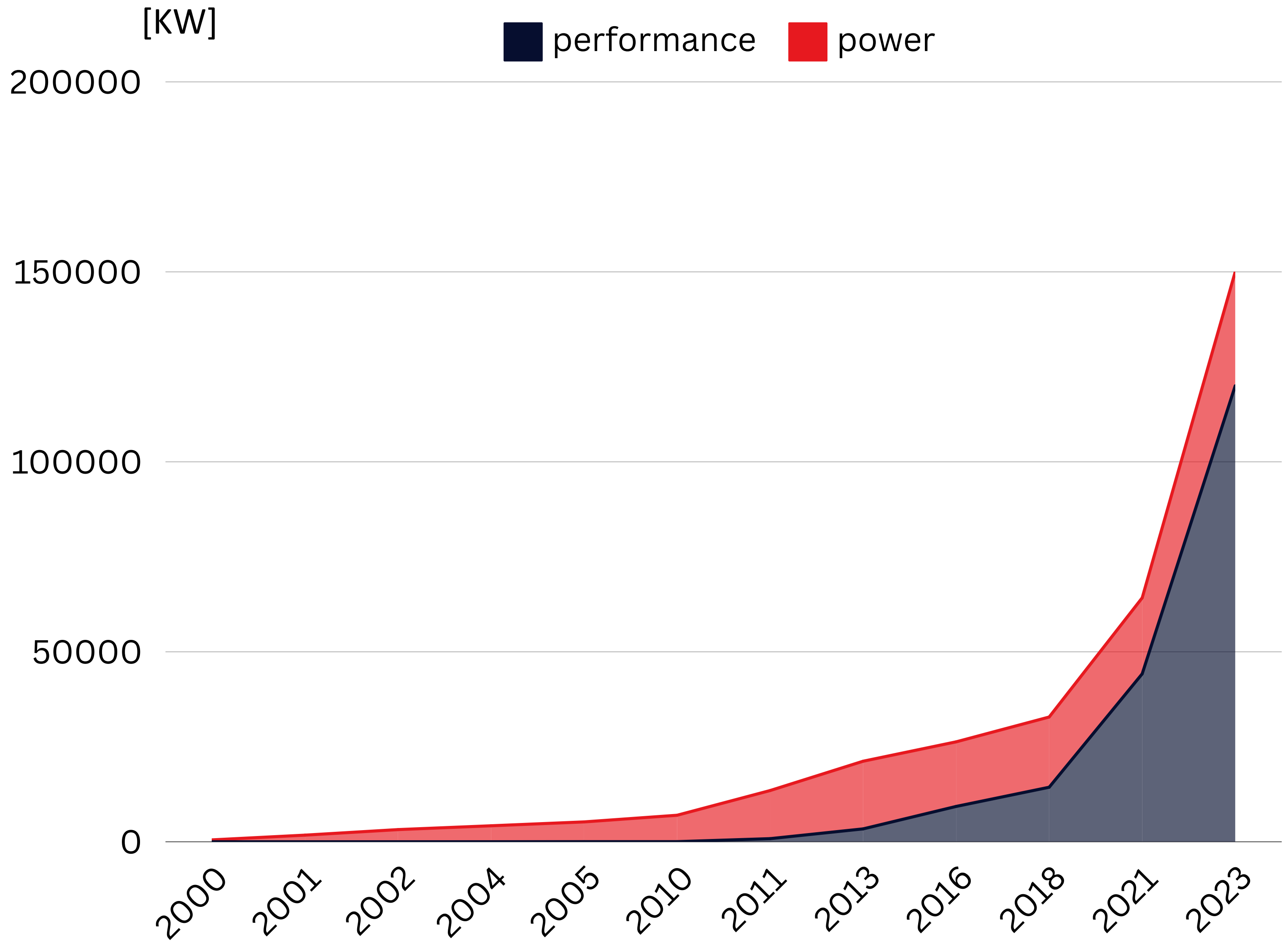


Applications : e.g,



/

Performance vs Energy consumption :



Top500 list performance and power data [2]

[1] blogs.vmware.com/apps/2018/09/vhpc-ra-part1

[2] www.top500.org/statistics/

Context

- HPC Systems exibit :

- Complex, interconnected, and with different specifications Hardware.
- Unpredictability in resource utilization due to the varying workloads.
- hardware and software failures.
- Applications change of phases.[1]
- varying system temperature.

require → **Online Monitoring & Dynamic Management**

- Some of dynamic management tools include :

Scheduling Algorithms, Autonomic Computing, and Control Theory Feedback [2]

---

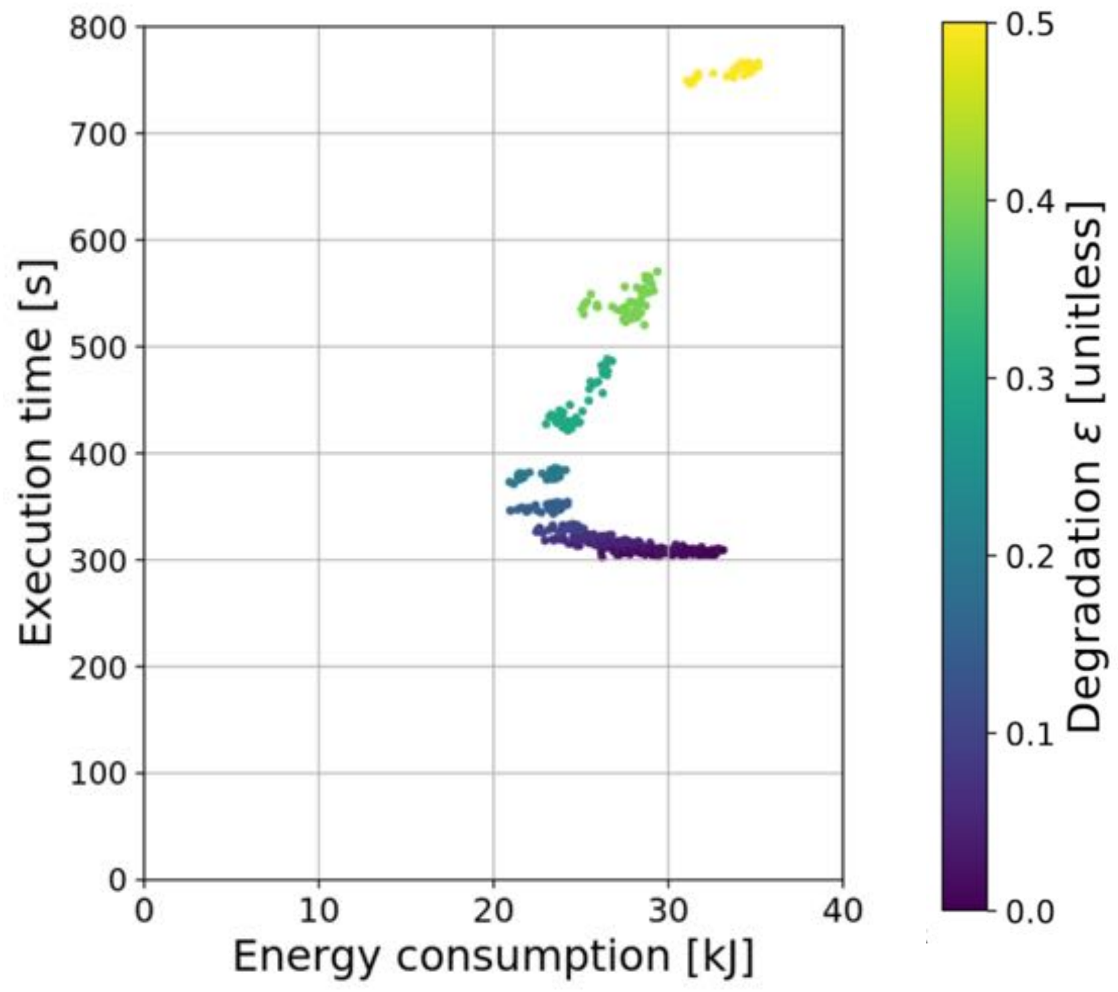
[1] S. Ramesh et al., "Understanding the Impact of Dynamic PowerCapping on Application Progress," in IPDPS, pp. 793–804, 2019.
[2] Joseph L. Hellerstein, Yixin Diao, Sujay Parekh, and Dawn M. Tilbury. 2004. Feedback Control of Computing Systems. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Context

- Global Objective :

- To apply a performance degradation on different benchmarks and study the tradeoff between the global benchmark execution time and energy consumption.

- Monitor and Control the Online Performance of the application using suitable sensors and control knobs

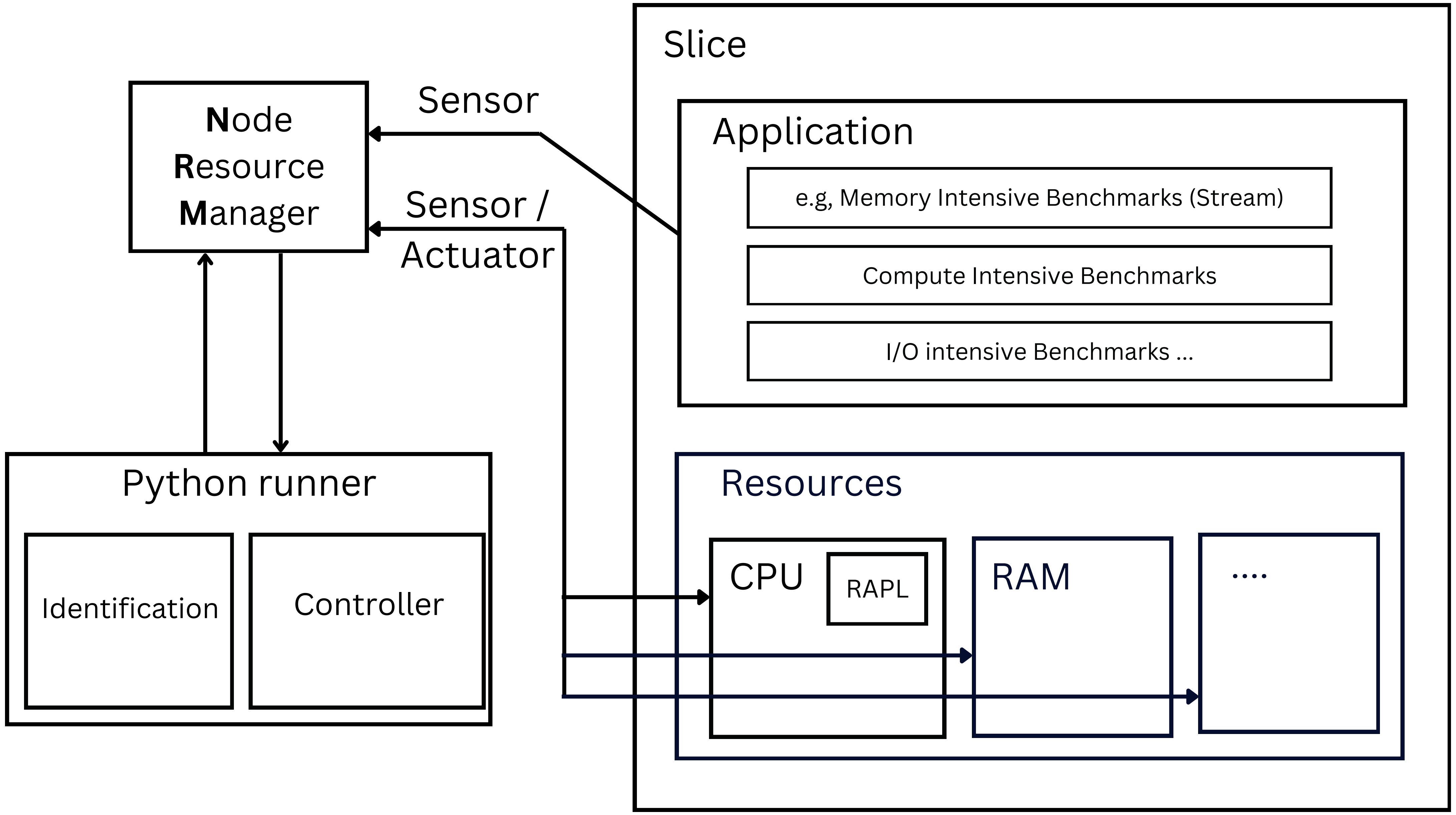


Execution time with respect to energy consumption. Color indicates the requested degradation level. Each point depicts a single execution[1]

[1] Sophie Cerf et al. “Sustaining Performance While Reducing Energy Consumption: A Control Theory Approach.” In: Euro-Par 2021: Parallel Processing.

System Architecture :

Software Stack : Argo Node Resource Manager Framework [1]



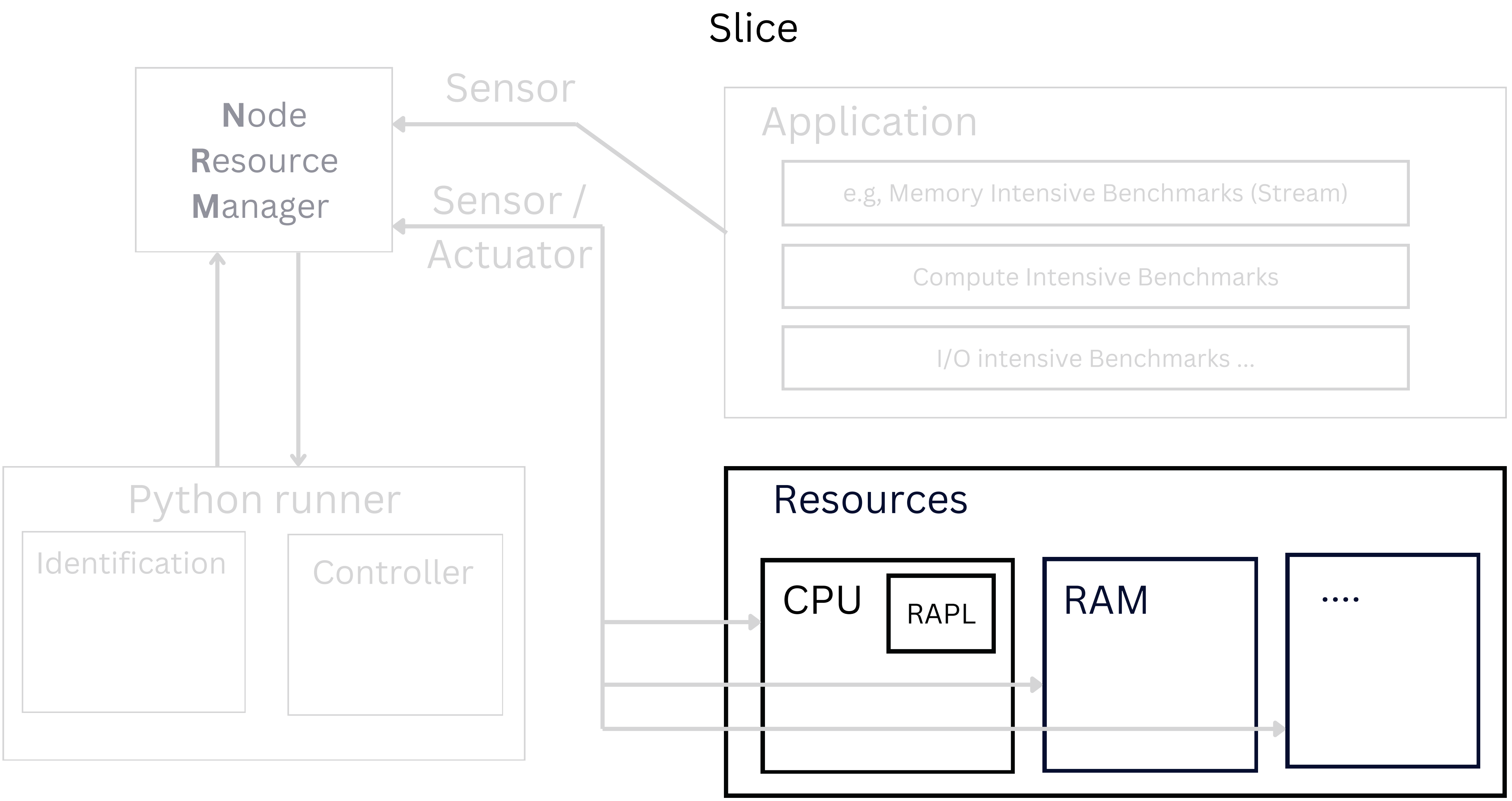
Platform : 1 Node from 3 different clusters of the Grid5000

[1] web.cels.anl.gov/projects/argo/overview/nrm/

System Architecture :

Software Stack : Argo Node Resource Manager Framework

- **The HPC system Hardware** is a single Node from three different clusters,equipped with powerful processors.

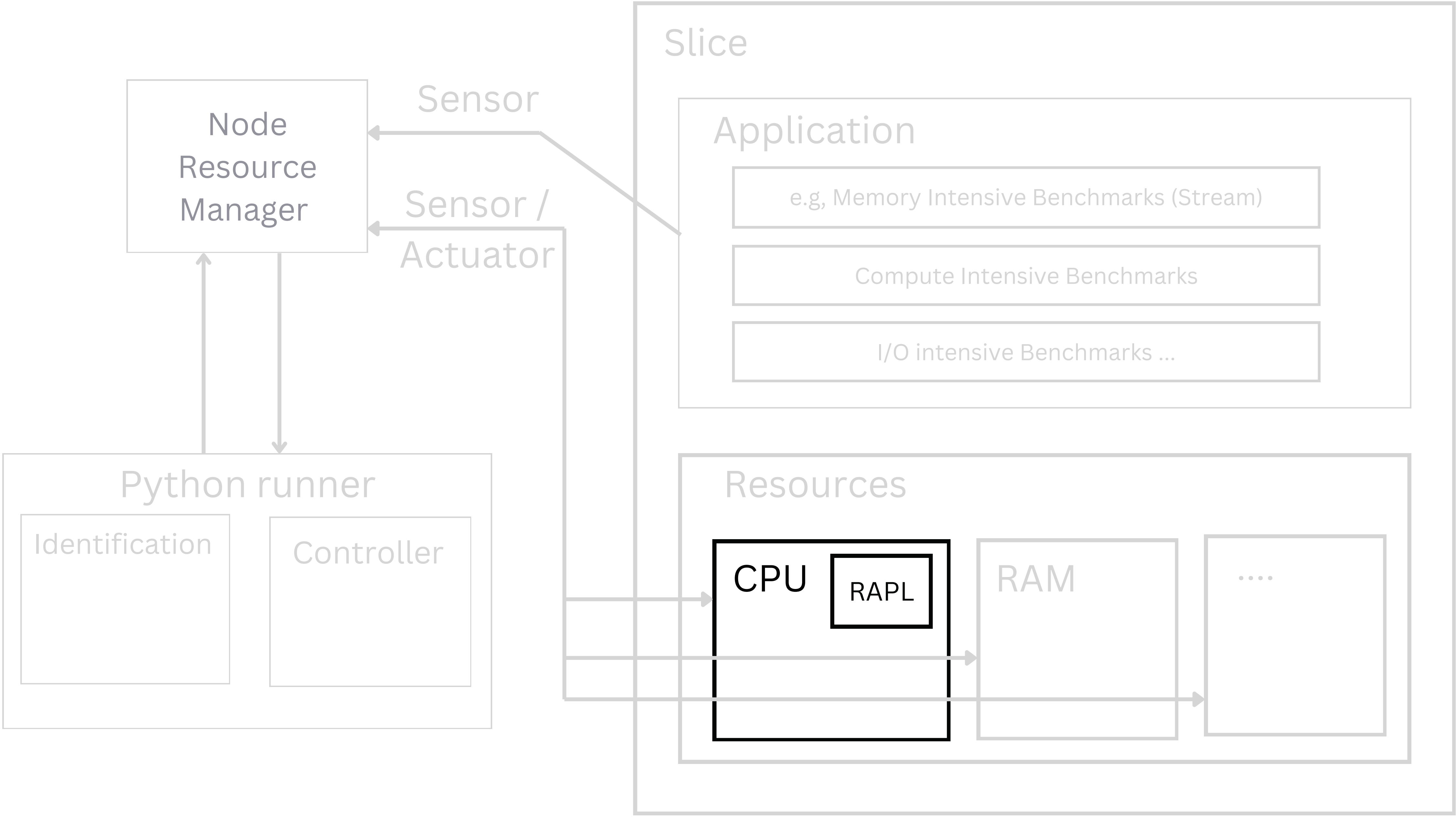


Platform : 1 Node from 3 clusters of the Grid5000

System Architecture :

- **RAPL** actuator which is an autonomous hardware solution implemented on Intel processors, it allows users to specify a power cap on the hardware. [1]

Software Stack : Argo Node Resource Manager Framework



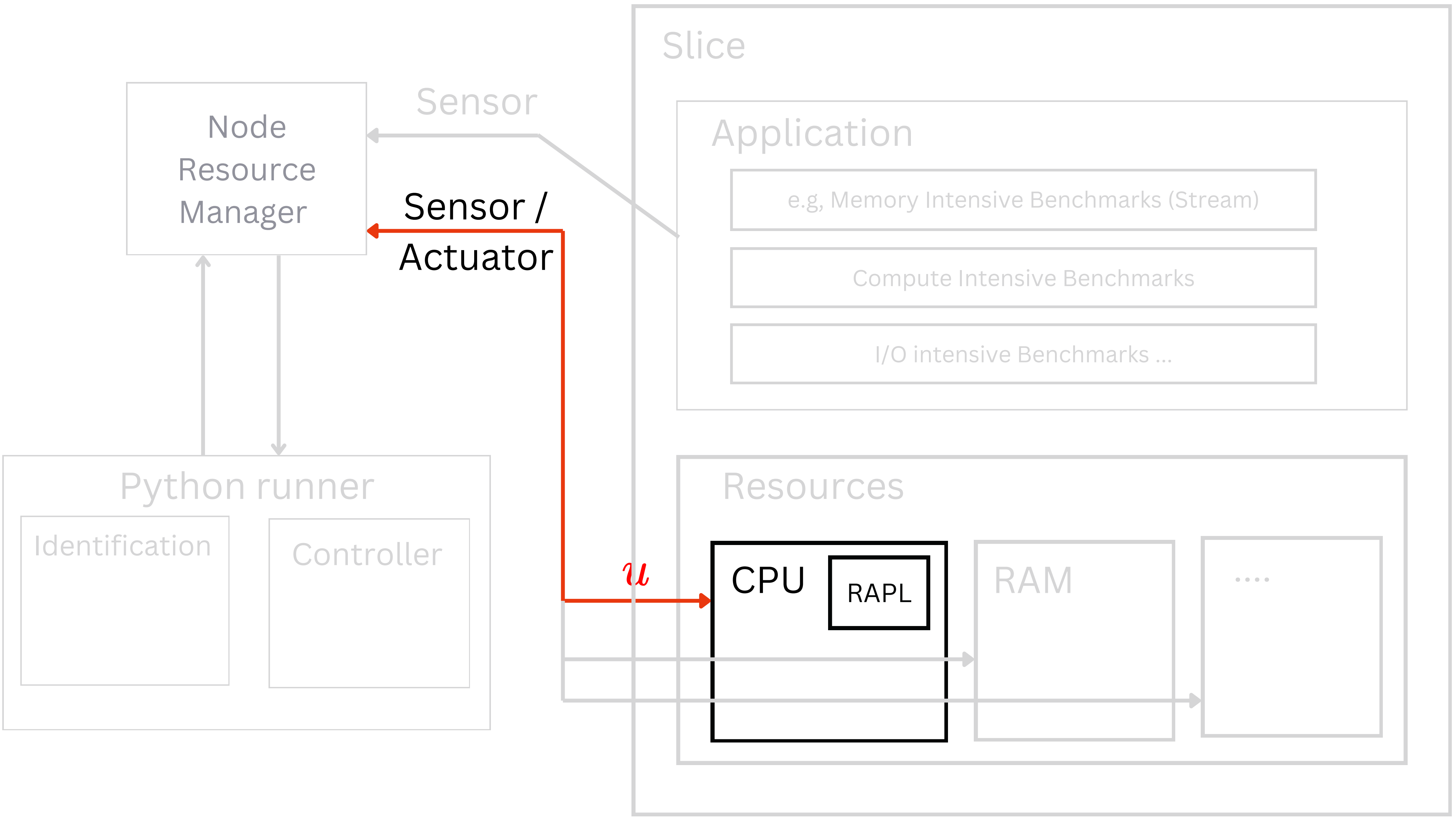
Platform : 1 Node from 3 clusters of the Grid5000

[1] David, H., et al.: RAPL: Memory Power Estimation and Capping. In: ISLPED. pp. 189–194. ACM (2010).

System Architecture :

Software Stack : Argo Node Resource Manager Framework

- Powercap : $u(t_i) = powercap(t_i)$

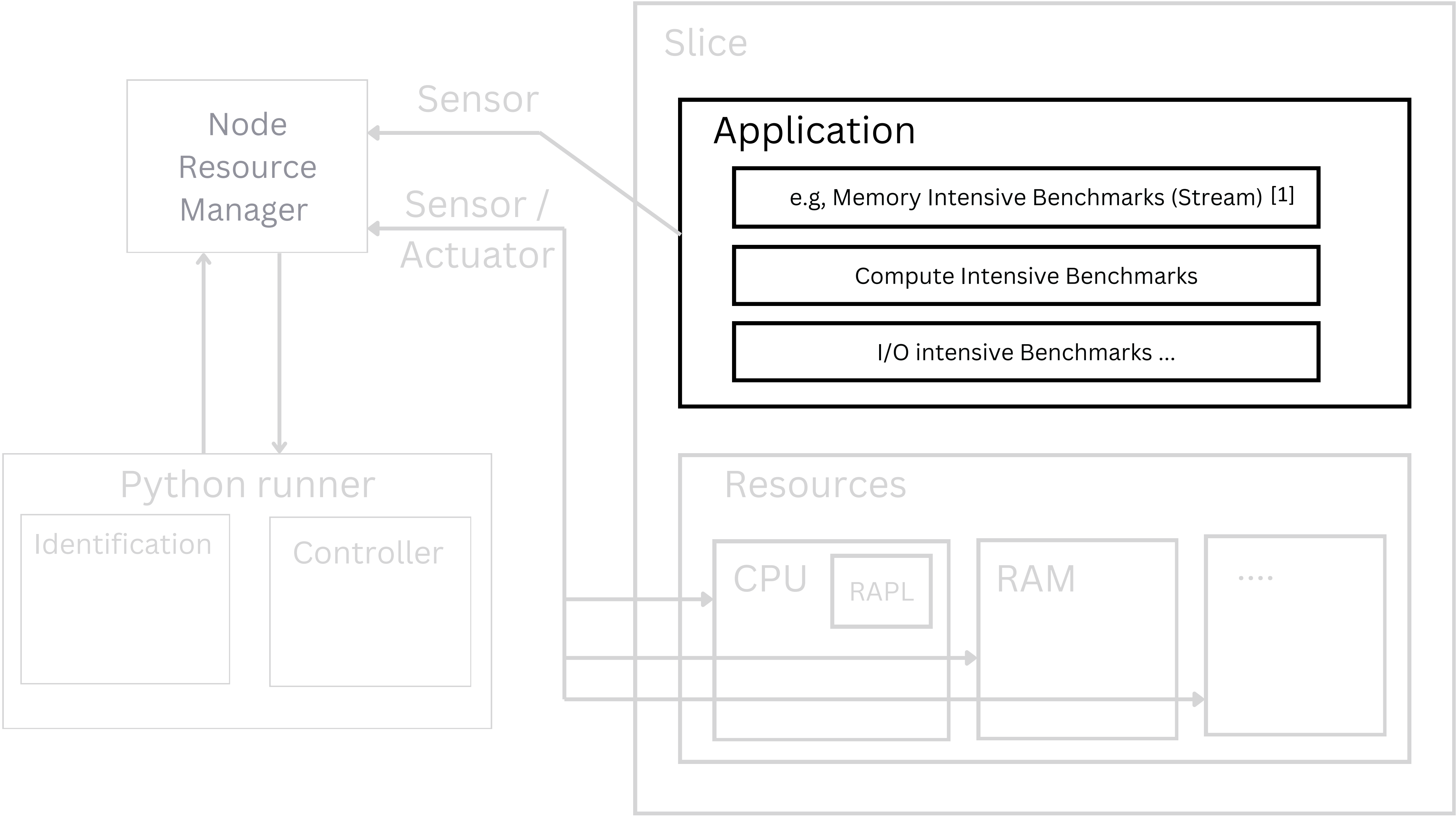


Platform : 1 Node from 3 clusters of the Grid5000

System Architecture :

- **The HPC application** used the Embarrassingly Parallel (EP) Compute intensive Benchmark.

Software Stack : Argo Node Resource Manager Framework



Platform : 1 Node from 3 clusters of the Grid5000

[1] Sophie Cerf et al. "Sustaining Performance While Reducing Energy Consumption: A Control Theory Approach." In: Euro-Par 2021: Parallel Processing.

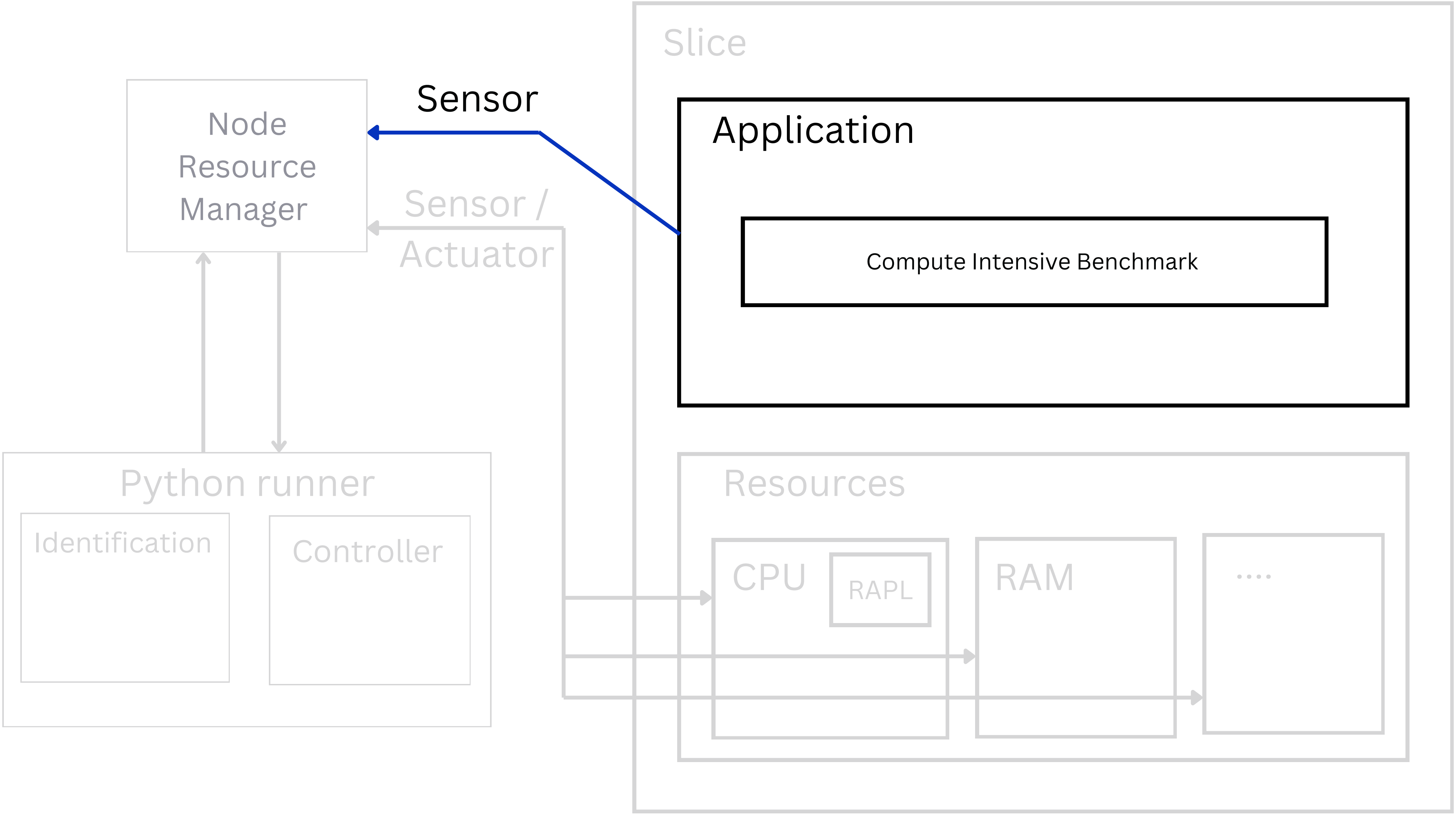
System Architecture :

Software Stack : Argo Node Resource Manager Framework

- Embeds a specialized library within the application, emitting "**heartbeats**" or messages at specific code points, indicating Application **progress.** [1]

- Application Progress:

$$y(t_i) = \underset{\forall k, t \in [t_{i-1}, t_i]}{\text{median}}\left(\frac{1}{t_k - t_{k-1}}\right)$$



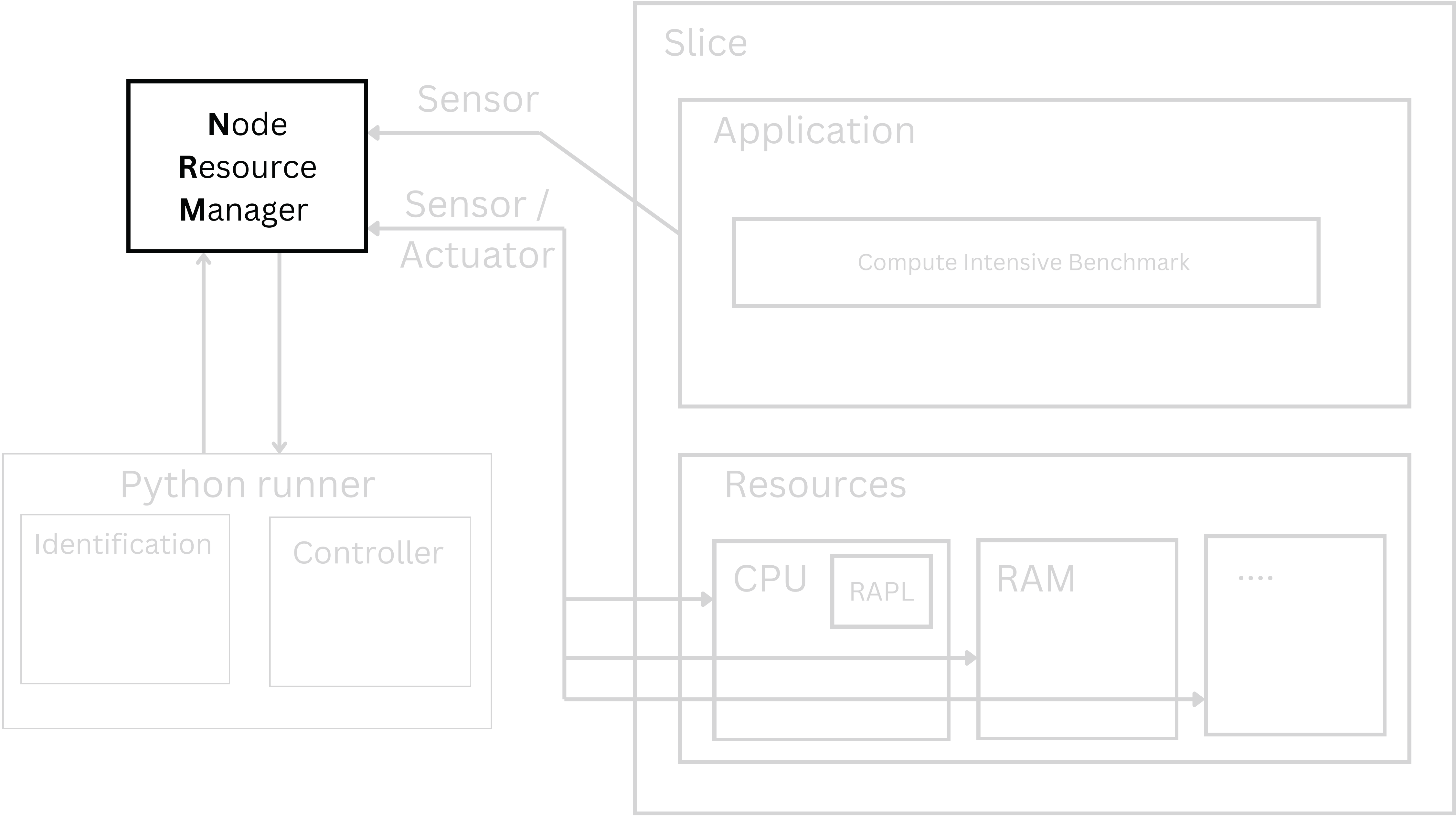
Platform : 1 Node from 3 clusters of the Grid5000

[1] S. Ramesh et al., "Understanding the Impact of Dynamic PowerCapping on Application Progress," in IPDPS, pp. 793–804, 2019.

System Architecture :

- **The Argo Node Resource Manager** acts as a central coordinator between the application and the underlying hardware, it is responsible for managing the tasks of sensing and control. [1]

Software Stack : Argo Node Resource Manager Framework



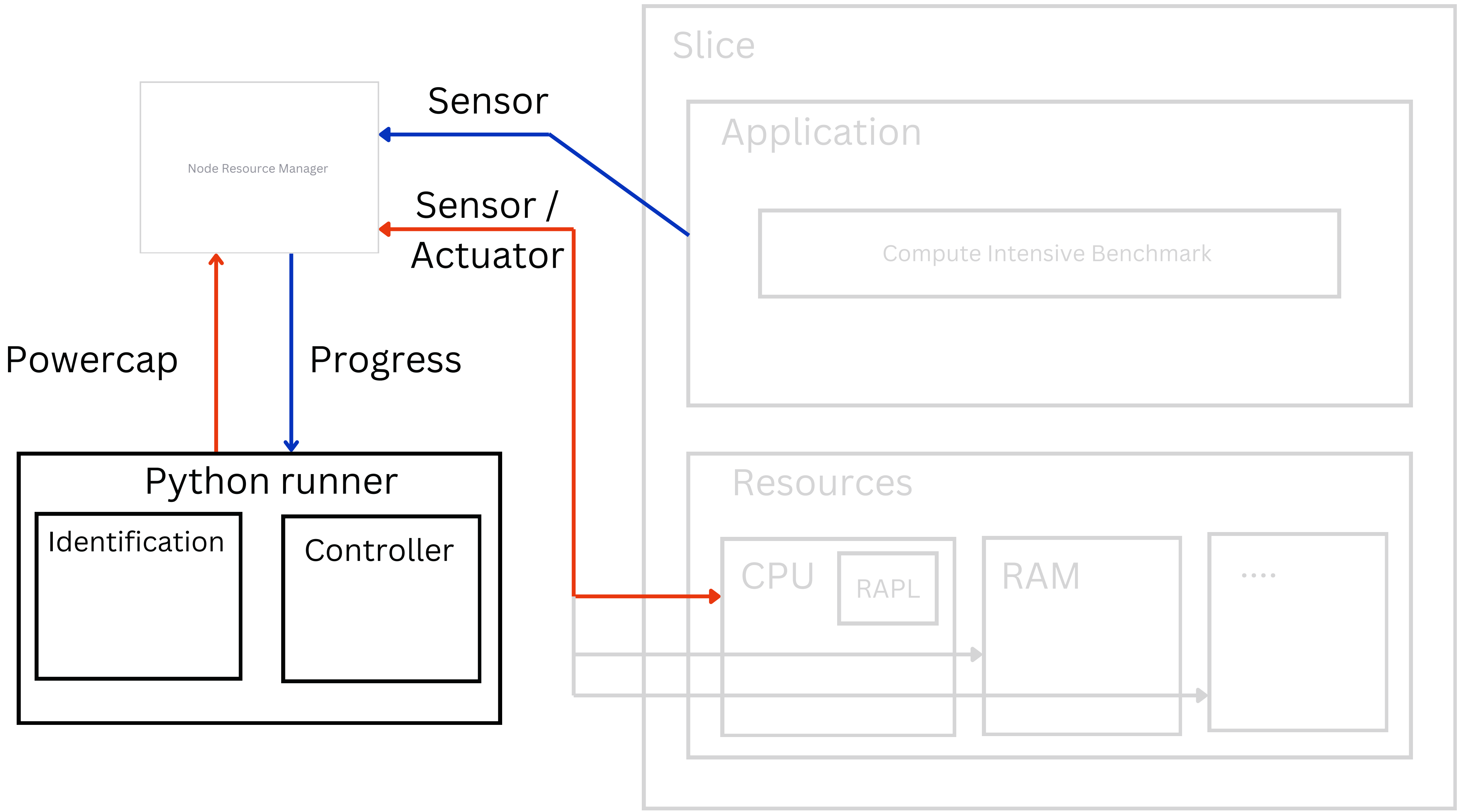
Platform : 1 Node from 3 clusters of the Grid5000

[1] web.cels.anl.gov/projects/argo/overview/nrm/

System Architecture :

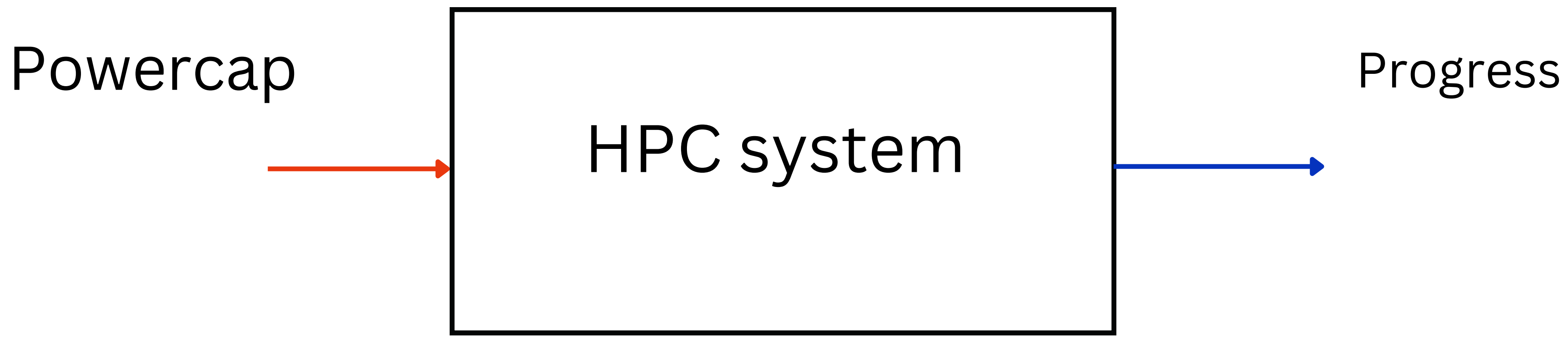
Software Stack : Argo Node Resource Manager Framework

- **The control loop** uses data from sensors to make informed decisions about power adjustments.

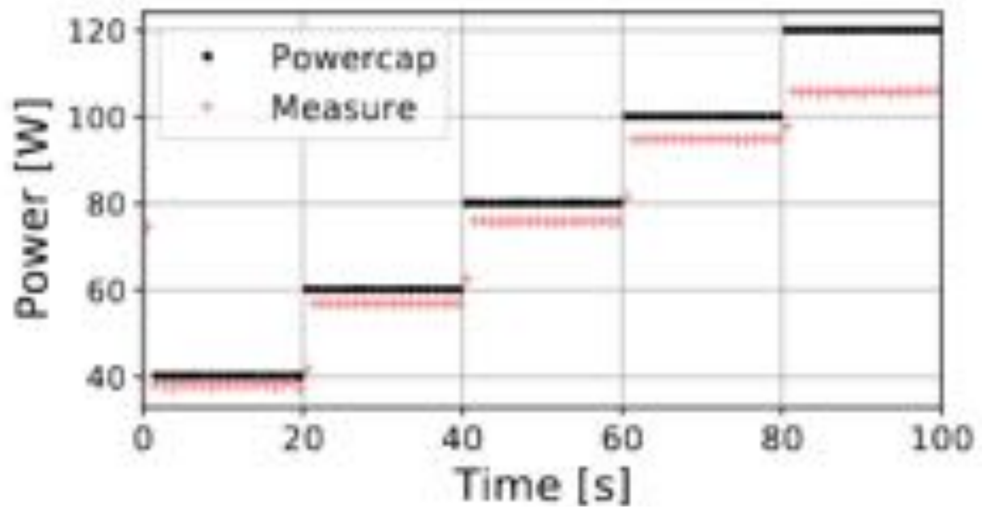


Platform : 1 Node from 3 clusters of the Grid5000

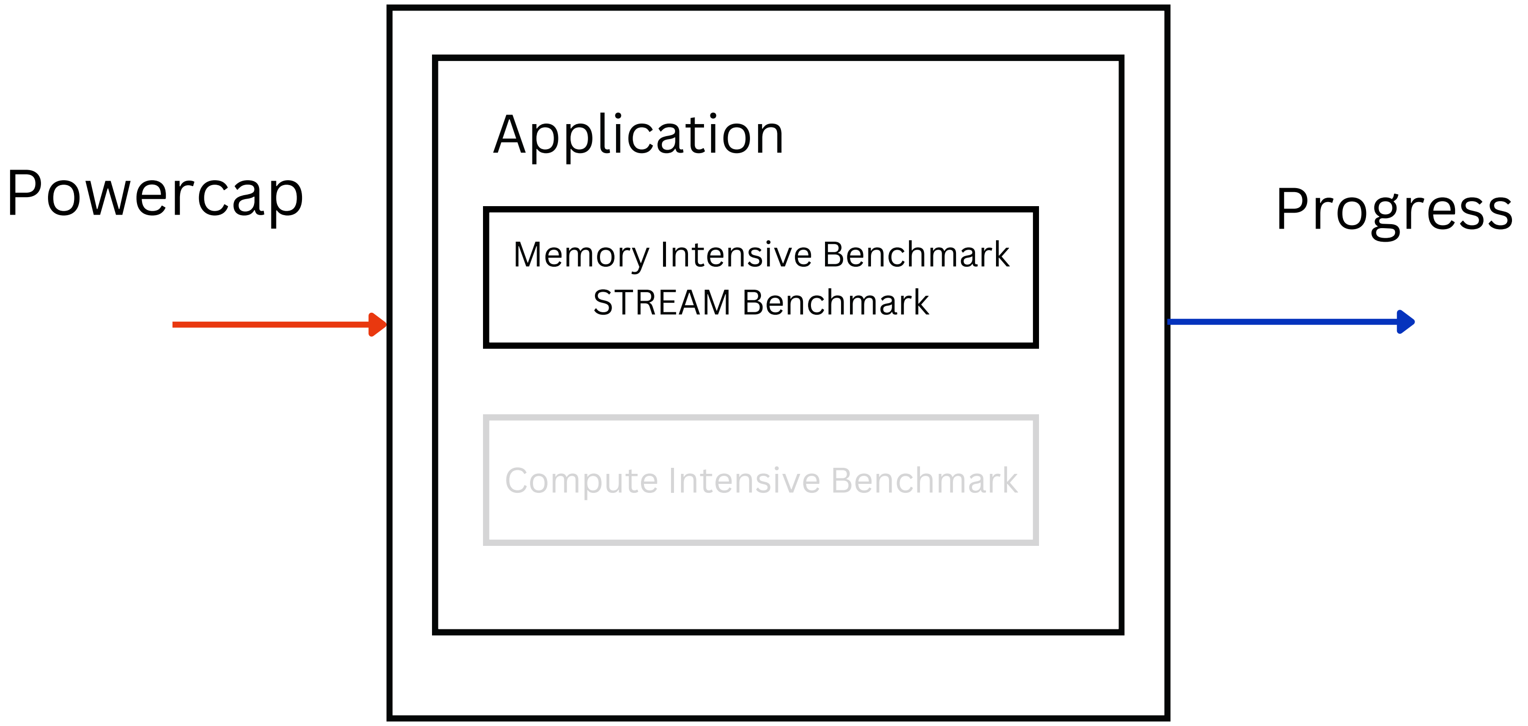
Approach and methodology



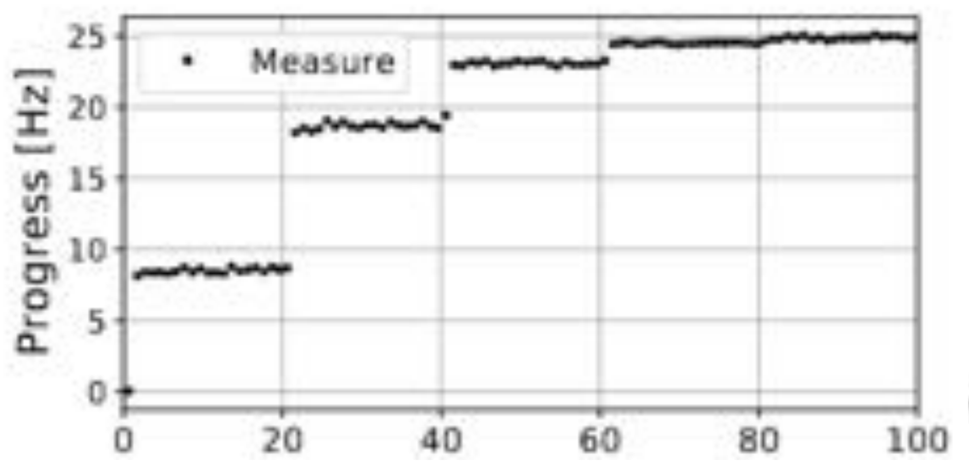
Approach and methodology



Gradual in RAPL powercap values from 40 to 120 [W]



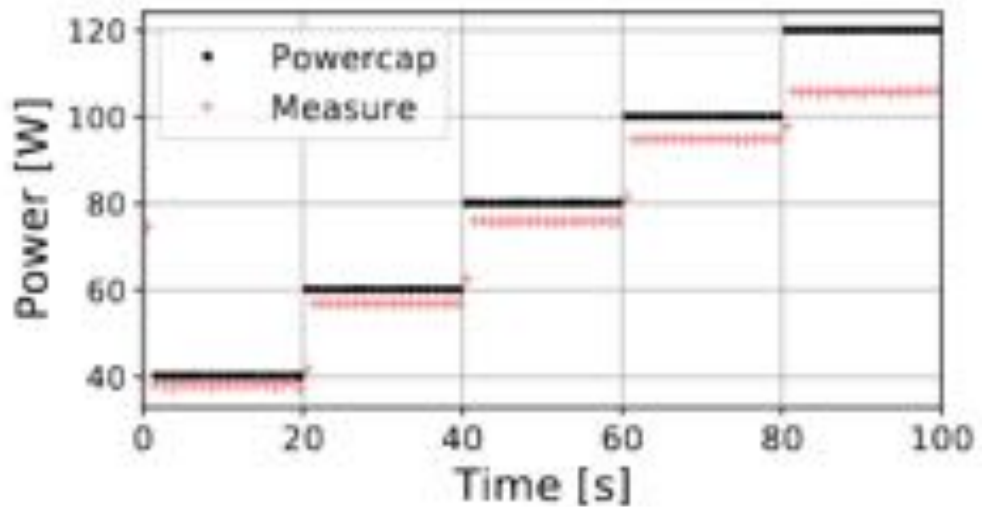
HPC system



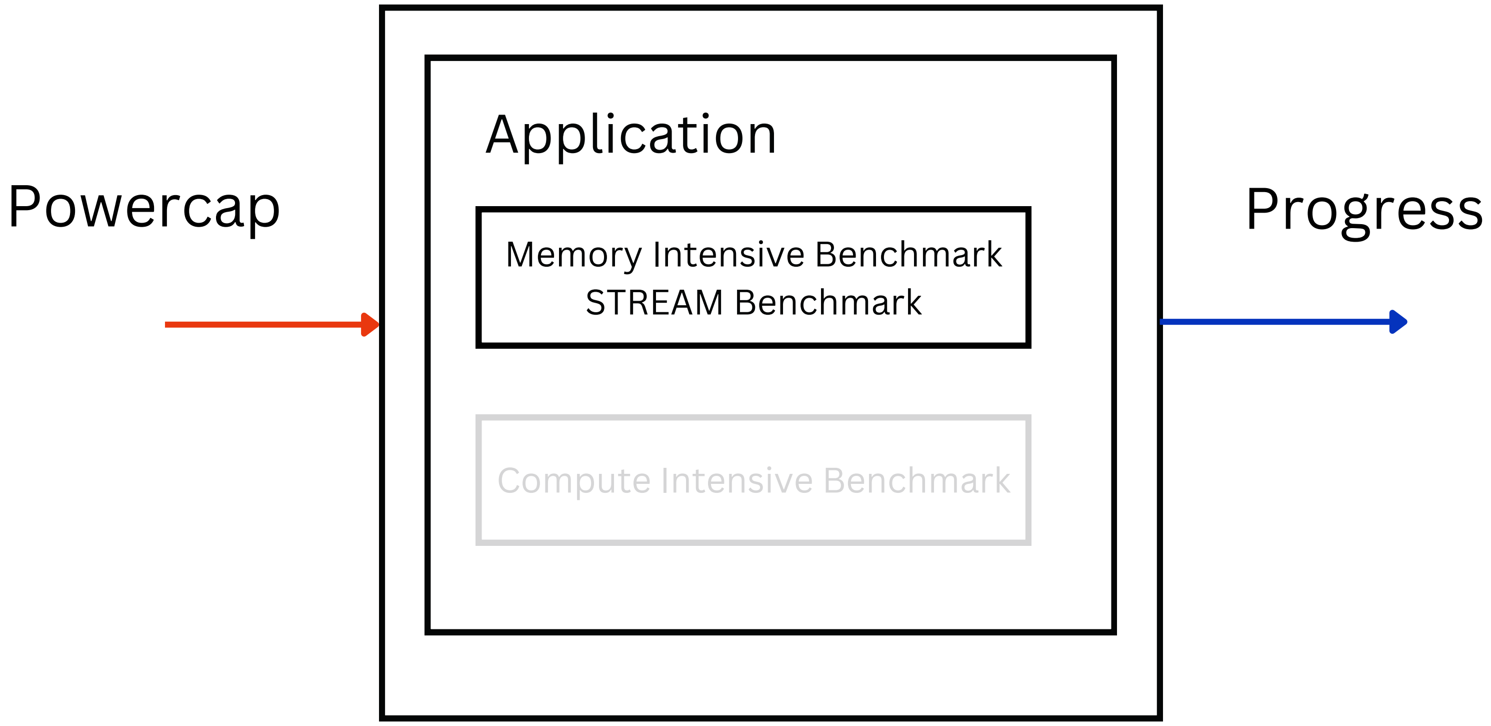
STREAM Benchmark Progress measure over time

[2] Sophie Cerf et al. "Sustaining Performance While Reducing Energy Consumption: A Control Theory Approach." In: Euro-Par 2021: Parallel Processing.

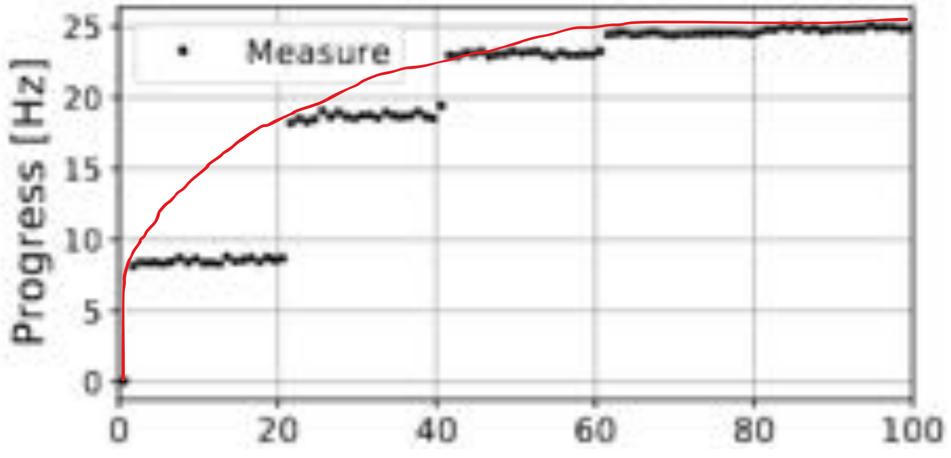
Approach and methodology



Gradual in RAPL powercap values from 40 to 120 [W]



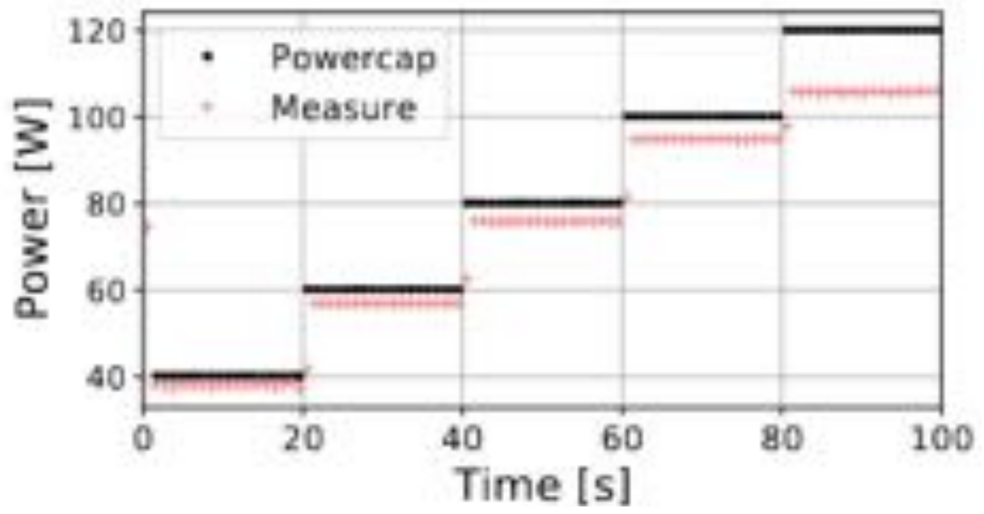
HPC system



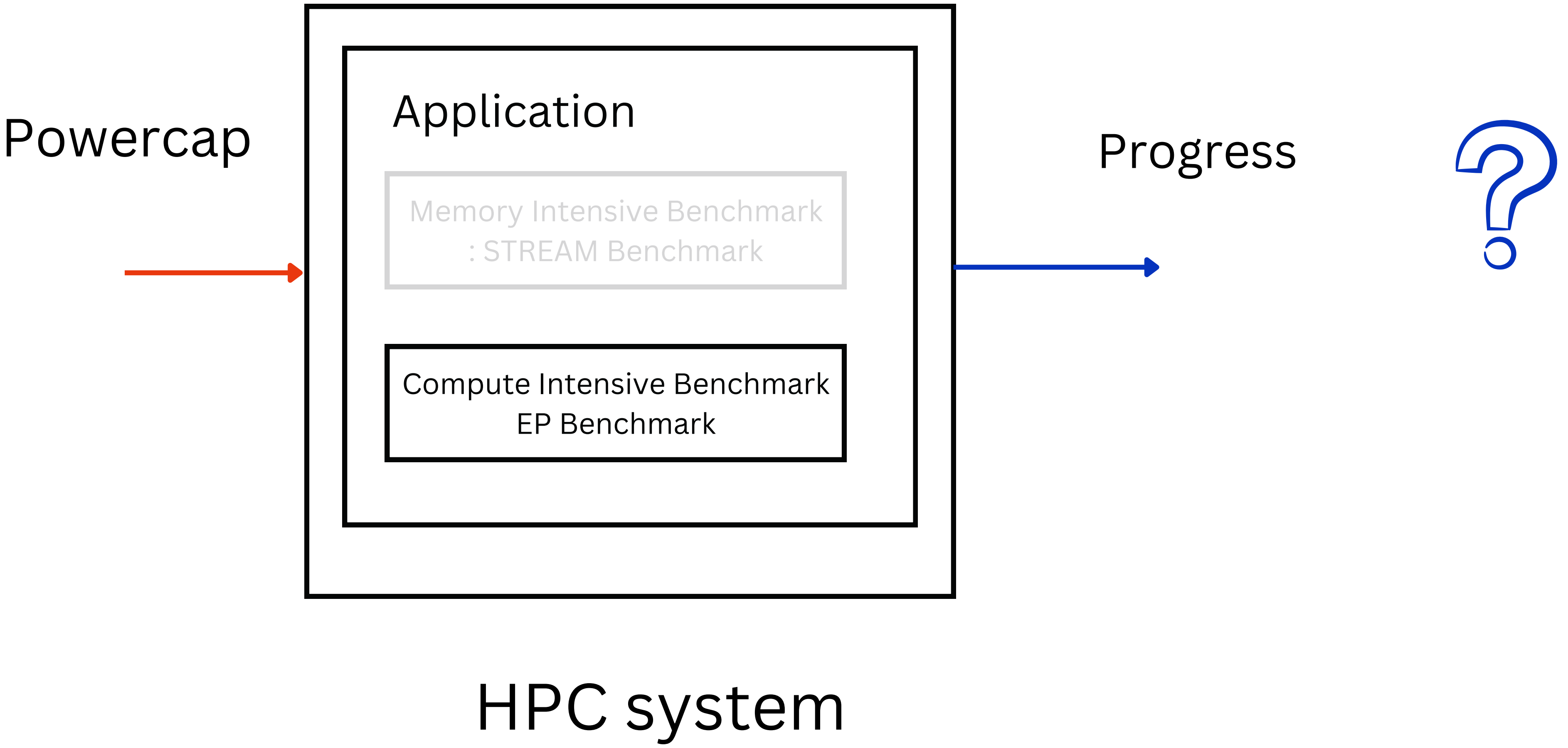
- The progress increases Exponentially which results in important energy savings over measurable performance degradation

[1] Sophie Cerf et al. "Sustaining Performance While Reducing Energy Consumption: A Control Theory Approach." In: Euro-Par 2021: Parallel Processing.

Approach and methodology



Gradual in RAPL powercap values from
40 to 120 [W]



Approach and methodology

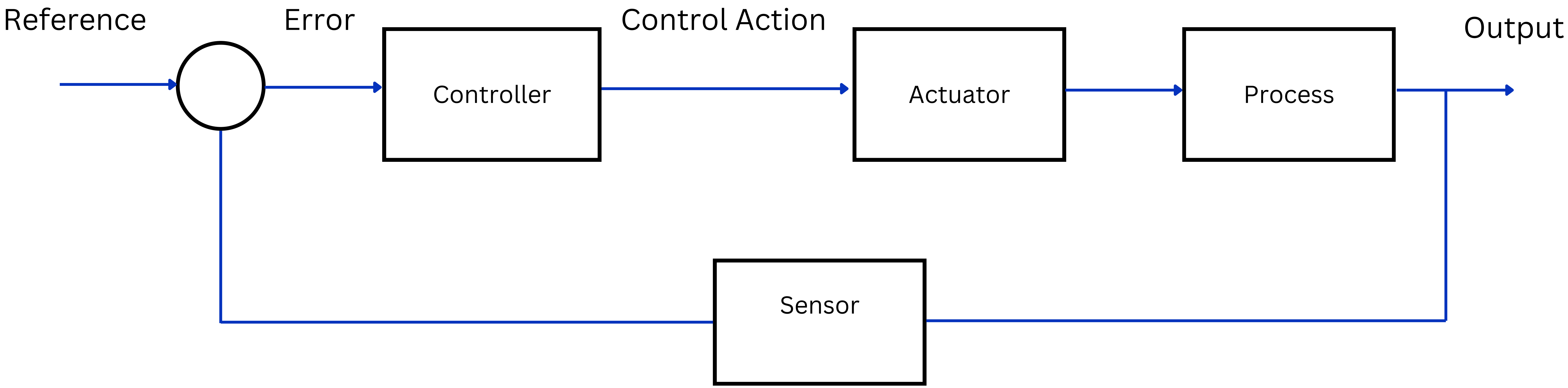
- **The objectives**

- To model the application and apply progressive powercaps to measure how much energy we can save by degrading the performance.

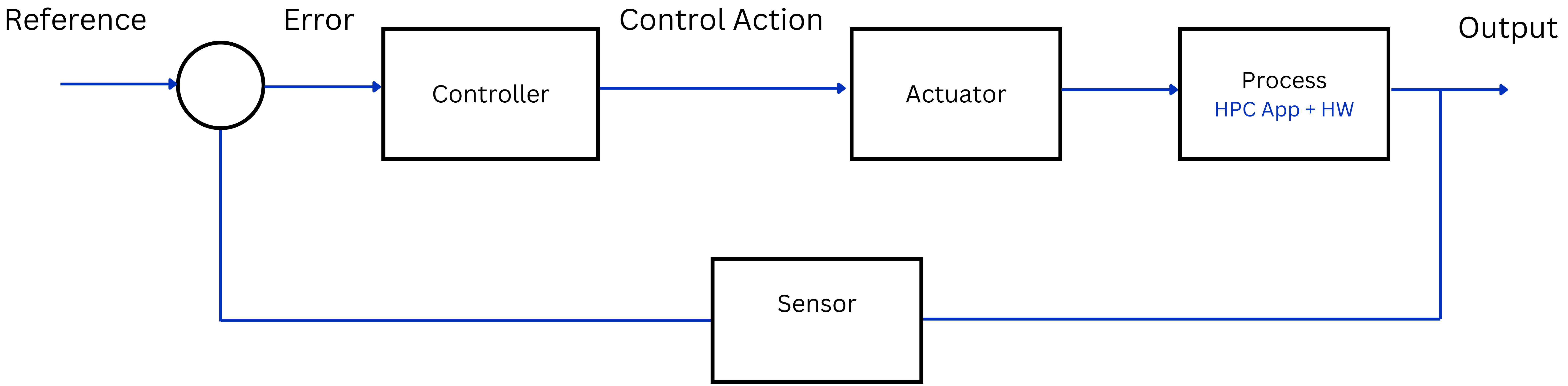
- Design a robust controller that monitors the application progress and measures the corresponding powercap.

- Use hierarchical control to correct RAPL inaccuracies.

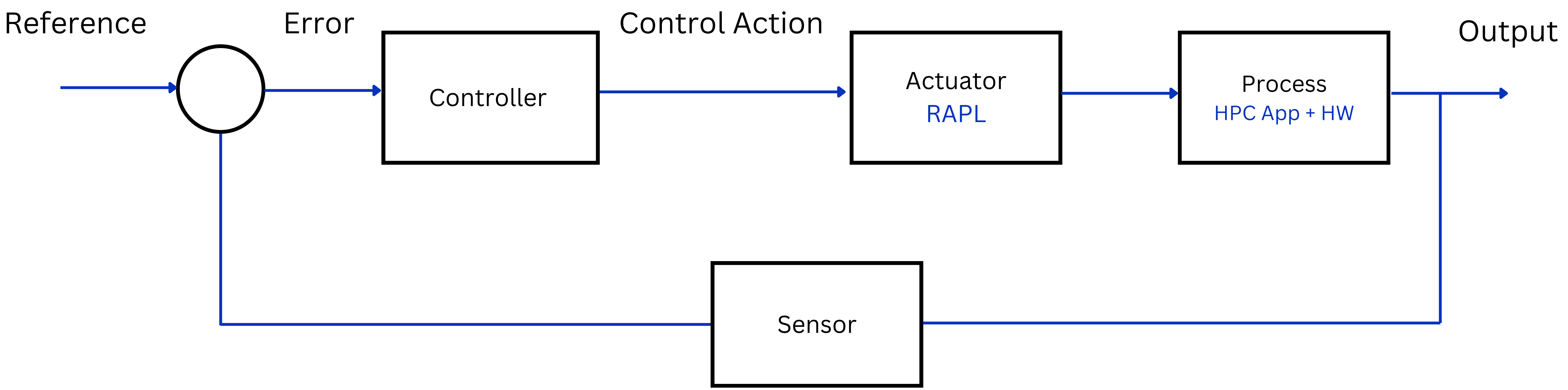
Control theory methodology



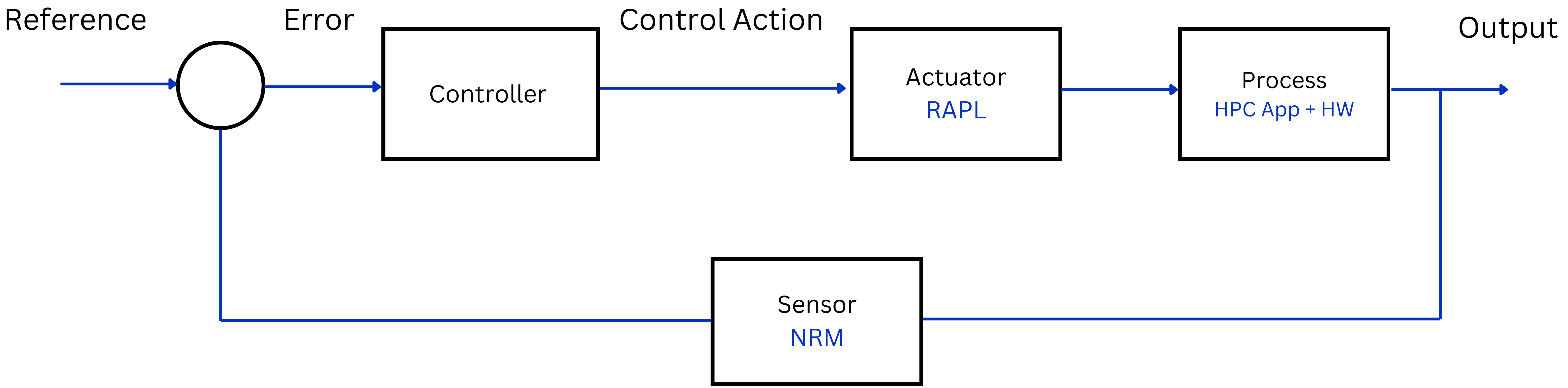
Control theory methodology



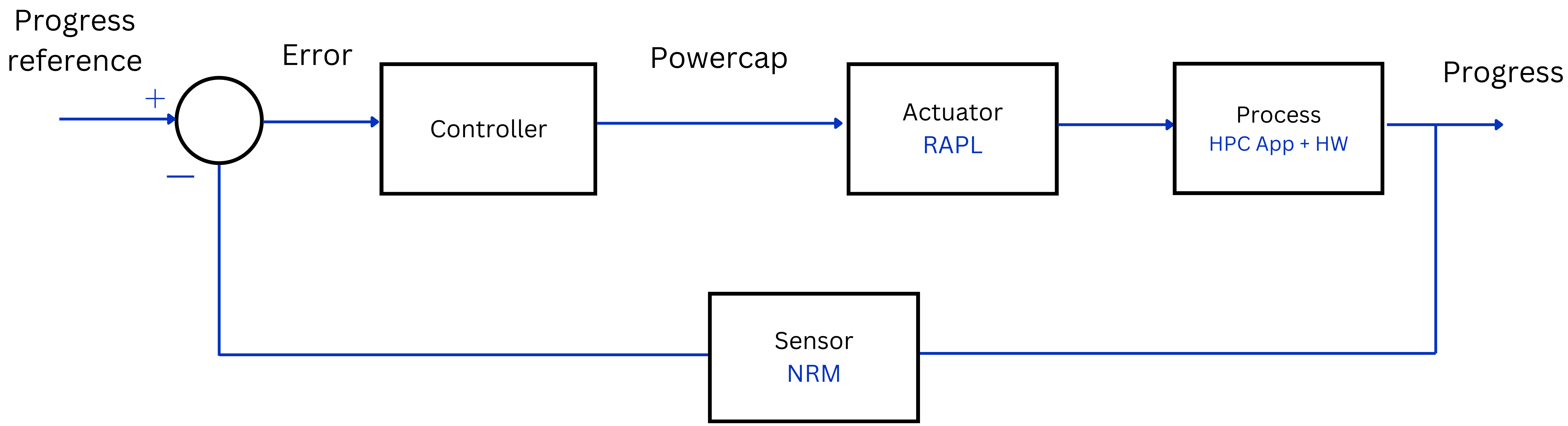
Control theory methodology



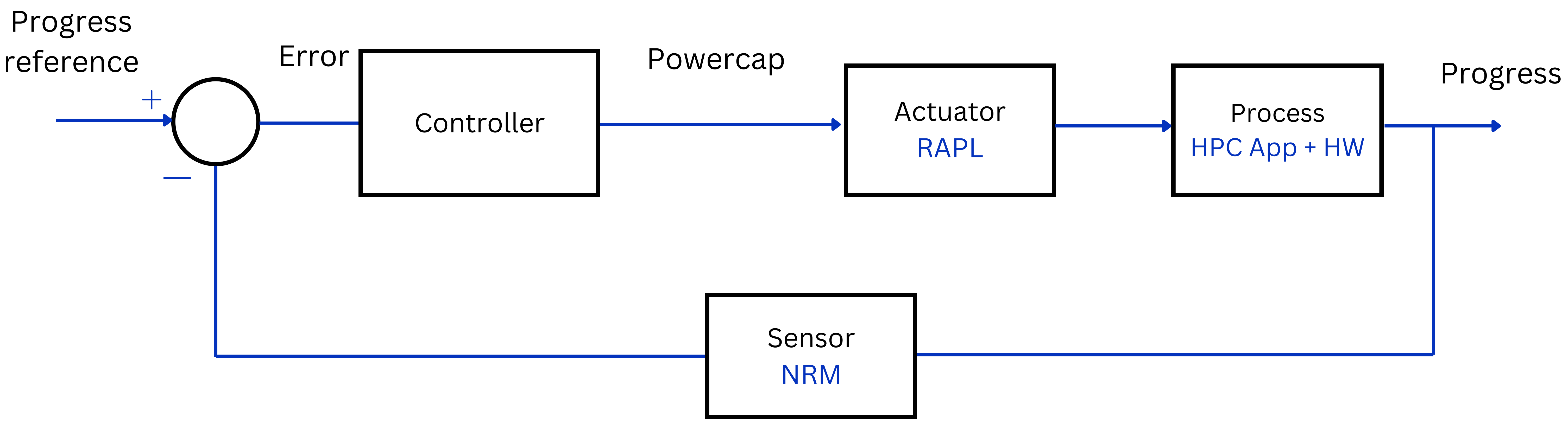
Control theory methodology



Control theory methodology



Control theory methodology



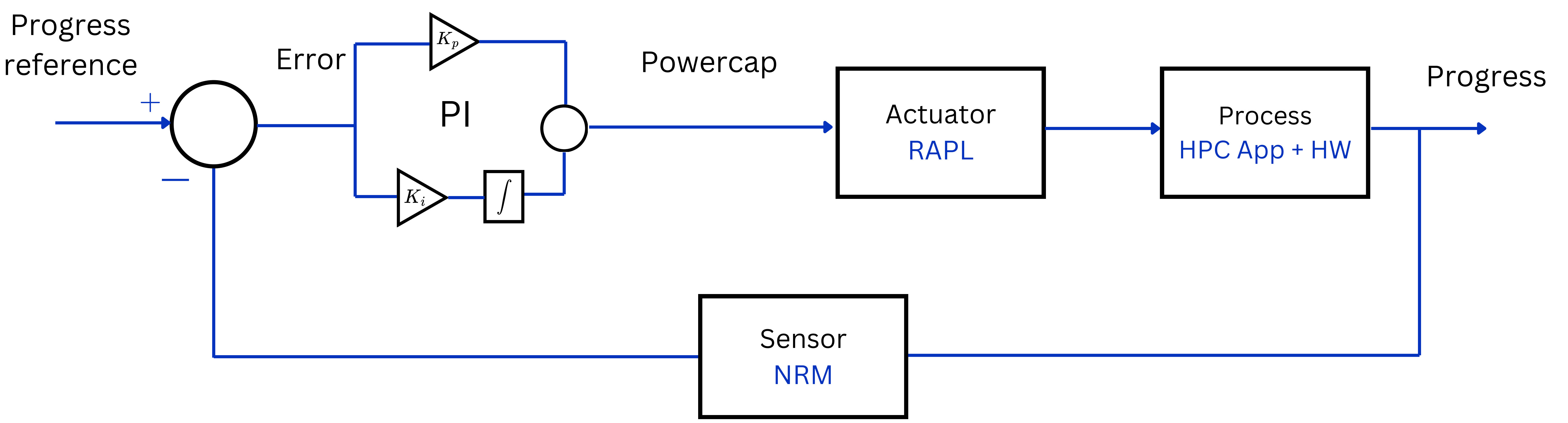
Degradation $\epsilon : 0 \leq \epsilon \leq 1$

Progress reference : $r(t) = (1 - \epsilon) \cdot y_{max}$

Progress : $y(t)$

Error : $e(t)$ = Progress reference - Progress

Control theory methodology



Degradation $\epsilon : 0 \leq \epsilon \leq 1$

Progress reference : $r(t) = (1 - \epsilon) \cdot y_{max}$

Progress : $y(t)$

Error : $e(t)$ = Progress reference - Progress

Powecap : $u(t) = K_p \cdot e(t) + K_i \cdot \sum_0^t e(i)$

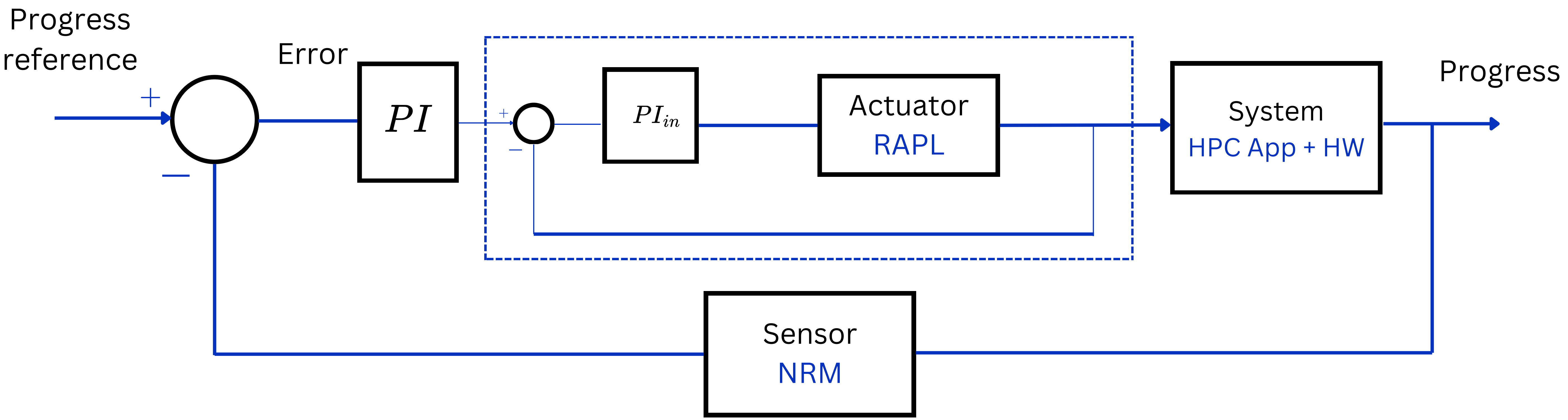
Control Objectives :

1 -> Application slowing down

- Ensure System Stability.
- Accurate reference tracking.
- Desired response characteristics : fast settling time, minimal overshoot, and smooth response to changes.
- robust PI tuning.

2 -> RAPL regulation

- internal fast control for RAPL regulation.



Modeling & Analysis :

1- RAPL Modeling

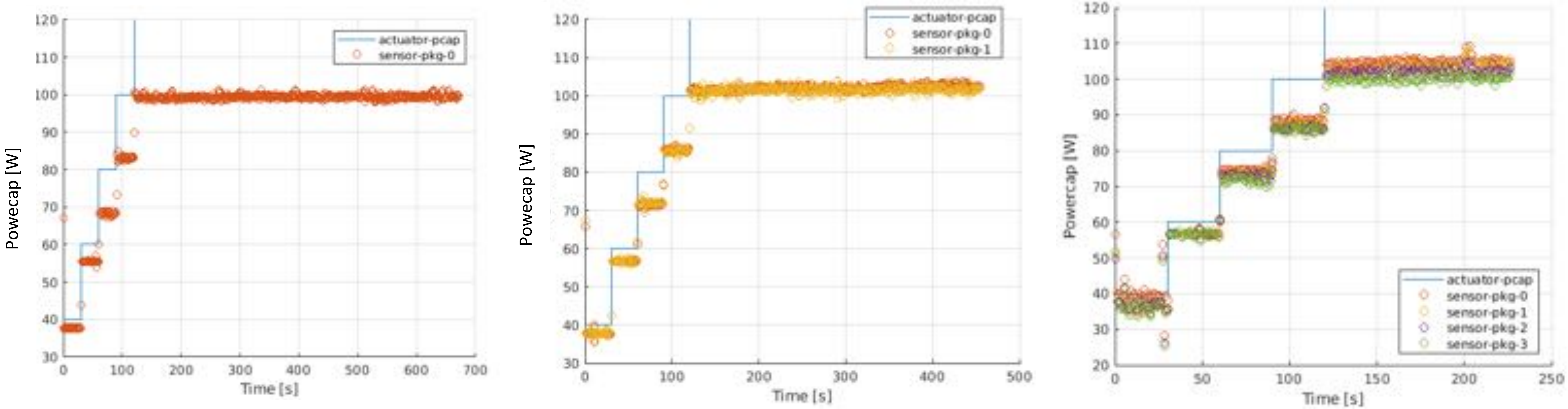
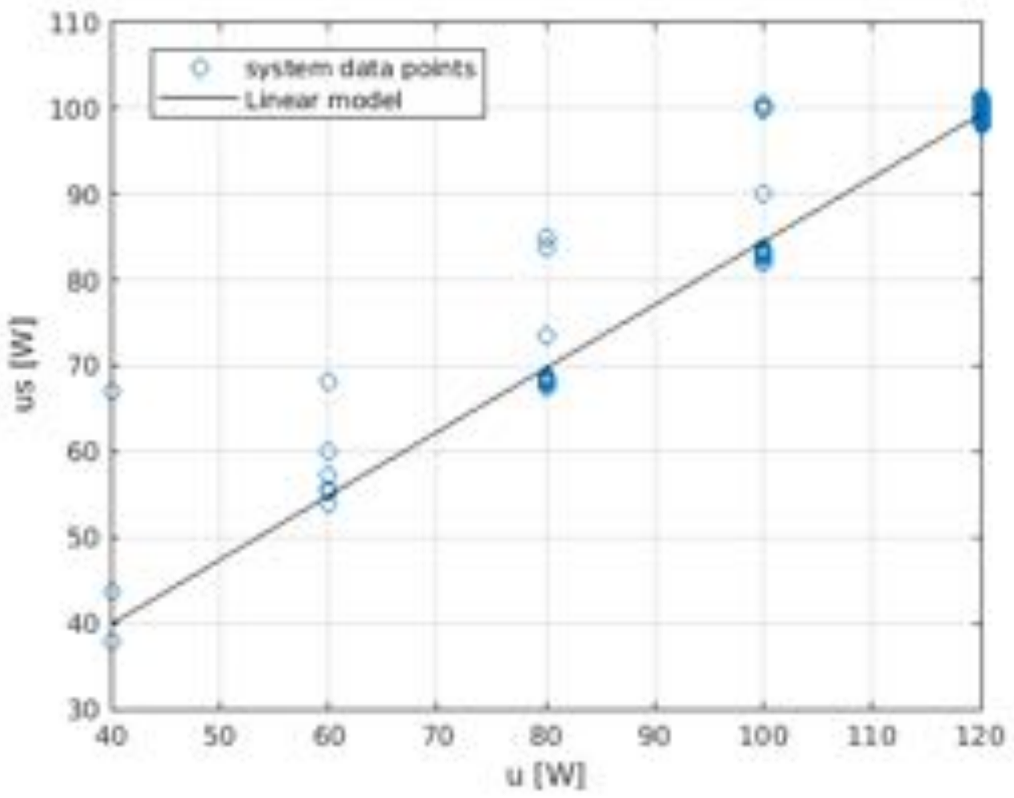


Fig. Requested and Measured powercap signals for Three Clusters (Gros, Dahu, Yeti)

Modeling :

1- RAPL Modeling



$$u = a \cdot u + b$$

- Where a and b are a constant parameters.

- Analysis :

- RAPL accuracy decreases linearly as the requested powercap increases.
- A linear Cluster and Time invariant 1^{st} order model is a good fit.

Modeling :

2- HPC Application Modeling

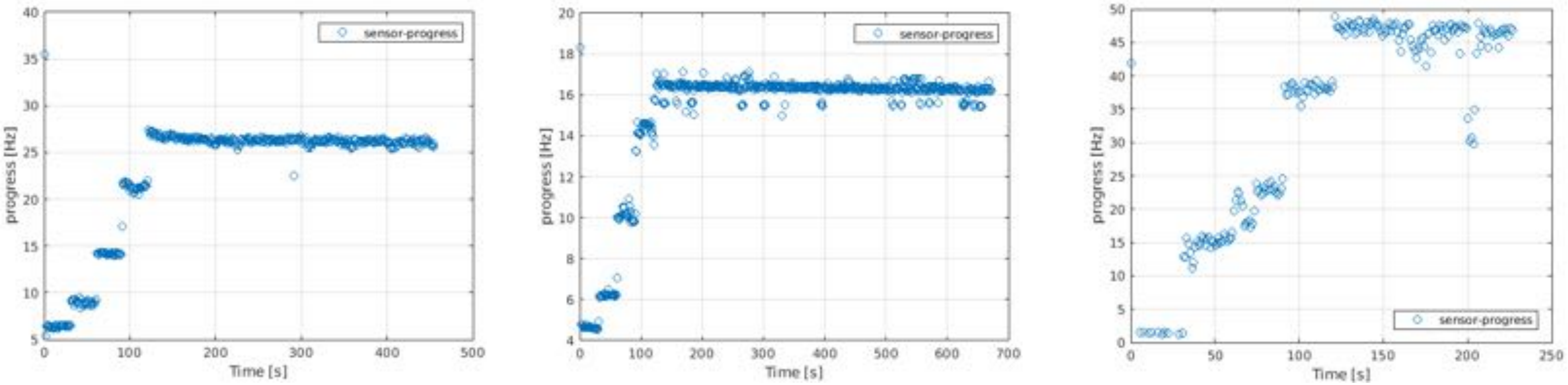
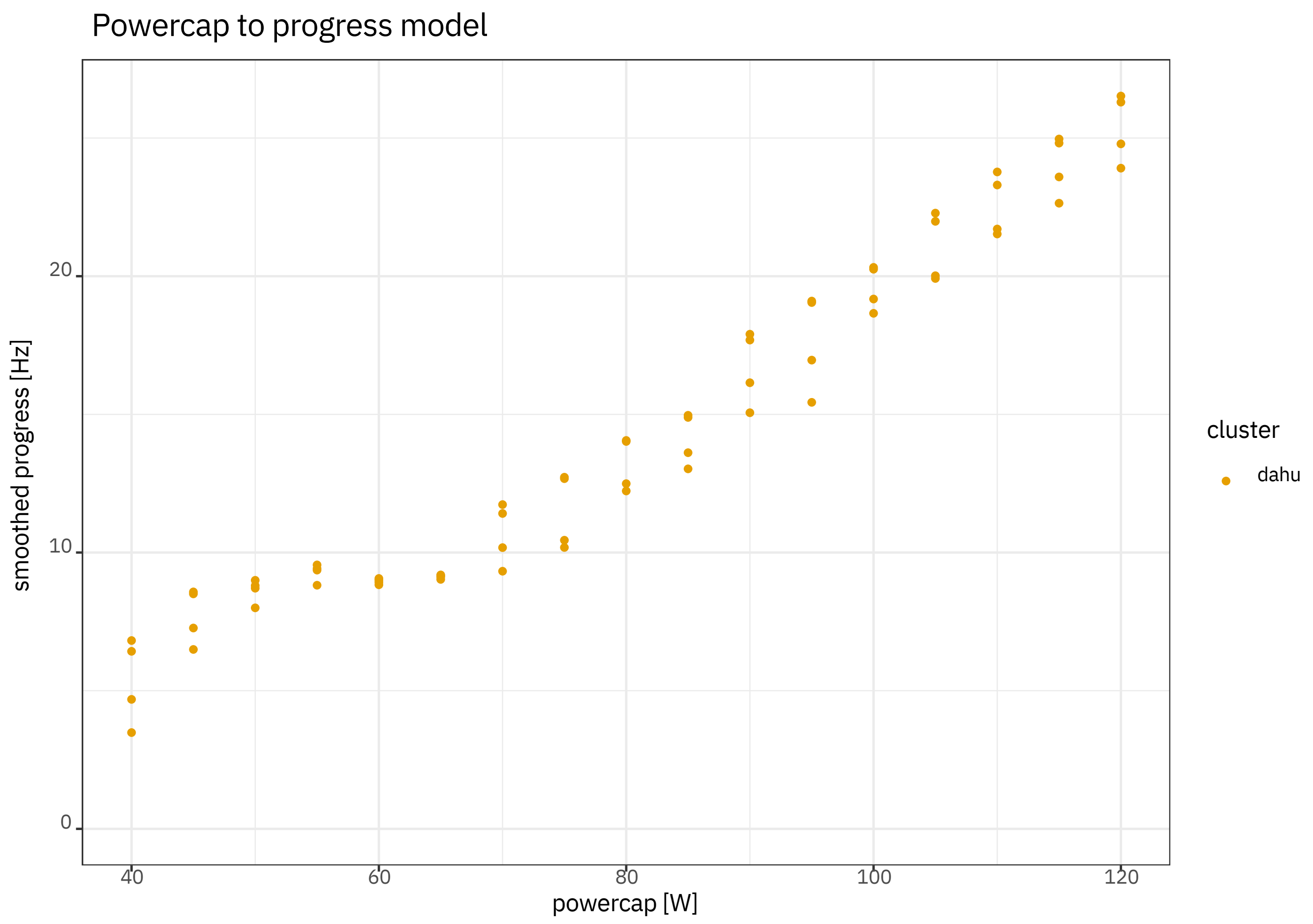


Fig. Application Progress Signals on Three Clusters

Modeling :

2- HPC Application Modeling



$$y = \phi + \alpha \cdot u$$

- Where ϕ and α are an unknown constant parameters.

- Analysis :

- The Static signal of Dahu Cluster shows a linear increase of the application progress with the increase of the powercap from 70 to 120 W.

- The System tend to be nonlinear for lower powercaps.

Results :

- Values and Analysis :

Degradation $\epsilon \approx 10\ \%$

PI inner loop gains : $K_{p_{in}} = 10, K_{i_{in}} = 25$

PI outer loop gains : $K_p = 0.2, K_i = 12$

Simulated Disturbances : band-limited white noise with a peak amplitude of 18 and a frequency of 0.1.

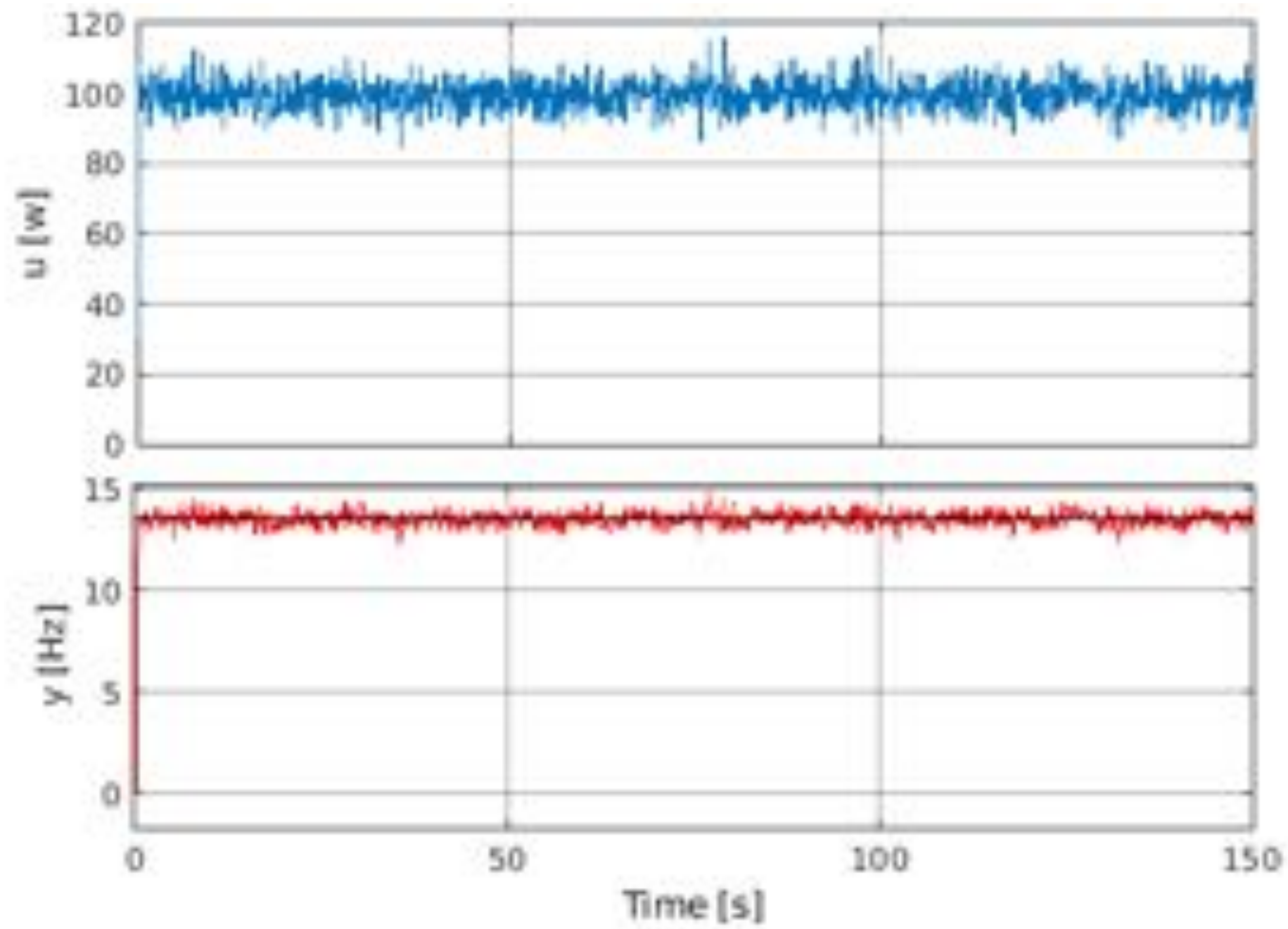


Fig. Controlled System Progress on Gros Cluster

Takeaways :

- Power management software should respond effectively to changes in application behavior (Application Phases, Memory, Network...)

- Control Theory is a strong and promising tool to regulate Computing Systems.

- Promising Results with +7% Execution time and -22% Energy Saving for Memory Intensive Applications (STREAM).[1]

- Expressing performance degradation as the primary design objective of the system would be interesting to Apply in certain use cases.

[1] Ismail Hawila, Sophie Cerf, Raphaël Bleuse, Swann Perarnau, Eric Rutten. Adaptive Power Control for Sober High-Performance Computing. CCTA 2022 - 6th IEEE Conference on Control Technology and Applications, Aug 2022, Trieste, Italy. pp.1-8.

References:

- Sophie Cerf et al. “Sustaining Performance While Reducing Energy Consumption: A Control Theory Approach.” In: Euro-Par 2021: Parallel Processing.

- Sophie Cerf et al. “Artifact and instructions to generate experimental results for the Euro-Par 2021 paper: ”Sustaining Performance While Reducing Energy Consumption: A Control Theory Approach”.”

- Argo Node Resource Manager. url: https://web.cels.anl.gov/projects/argo/overview/nrm/ (visited on 07/24/2023).

- S. Ramesh et al., “Understanding the Impact of Dynamic Power Capping on Application Progress,” in IPDPS, pp. 793–804, 2019.

Appendix :

1 - Table of clusters Hardware characteristics :

Cluster	Nodes	Sockets	CPU	Cores/CPU	Memory
gros	124	1	Intel Xeon Gold 5220	18	96 GiB
dahu	32	2	Intel Xeon Gold 6130	16	192 GiB
yeti	4	4	Intel Xeon Gold 6130	16	768 GiB