An interoperable runtime for distributed power management of large-scale HPC systems based on DDS.

Giacomo Madella

Alma Mater Studiorum, Università di Bologna
Research Topic:? HPC?

1 Introduction

In recent years, the rapid growth of super-computing systems has led to an increased demand for efficient power management strategies. As the landscape of computational capabilities continues its rapid expansion, the imperative for sustainable power management solutions becomes even more pronounced. The formidable computational power wielded by modern supercomputers is often juxtaposed with the considerable energy consumption required to fuel their operations. In this context, the pursuit of innovative power management strategies becomes not only an economic consideration but a pivotal endeavor to curtail the ecological ramifications of escalating energy consumption.

Given the intricate interplay between power efficiency, computational performance, and environmental responsibility, the development of advanced power management frameworks assumes paramount significance. It is within this framework that the present research aims to contribute. By harnessing the capabilities of open-source technologies and leveraging the efficiency-enhancing potential of Data Distribution Service (DDS) and Real-Time Publish-Subscribe (RTPS), we endeavor to engineer a dynamic and responsive power management runtime for High-Performance Computing (HPC) environments.

In light of the existing proprietary solutions that are often tailor-made for specific computing centers, the pursuit of open-source alternatives gains traction. Open-source power management solutions not only offer adaptability to the unique demands of different HPC ecosystems but also foster collaborative innovation across the broader scientific community.

In the subsequent sections of this research, we will delve deeper into the intricacies of designing and implementing our proposed runtime power management system. By integrating the efficiencies of DDS and drawing inspiration from successful open-source models, we aim to contribute to the evolving landscape of power management for HPC. Through this endeavor, we aspire to not only optimize computational efficiency and resource utilization but also to cultivate a greener and more sustainable future for supercomputing systems.

1.1 DDS & RTPS

DDS (Data Distribution Service) and RTPS (Real-Time Publish-Subscribe) constitute two pivotal technologies in the realm of distributed and real-time communications. These technologies play a critical role in enabling efficient and reliable data transmission among interconnected devices and applications, holding particular significance in intricate scenarios such as embedded systems, the Internet of Things (IoT), and high-performance applications like High-Performance Computing (HPC).

Specifically, DDS serves as a distributed communication framework that facilitates data exchange among software components distributed across heterogeneous networks. Built upon a publish-subscribe model, DDS establishes a mechanism for efficient data sharing. On the other hand, RTPS serves as the underlying protocol employed by DDS to realize the publish-subscribe paradigm within real-time networks. RTPS focuses on the dependable delivery of real-time messages, ensuring that data reaches the appropriate recipients in the most efficient manner. This protocol manages critical aspects such as data flow control, node synchronization, and quality of service management.

2 State of the Art

The state of the art can be delineated into two key realms pertinent to the research study. Firstly, a comprehensive overview of the existing landscape in HPC power management reveals noteworthy developments and challenges that lay the groundwork for the pursuit of interoperable solutions. Additionally, an exploration of the domain of Data Distribution Service (DDS)

2.1 HPC Power Management

In the landscape of power management within High-Performance Computing exhibits a dichotomy marked by non-interoperable solutions and various components interwoven within the HPC powerstack. Despite the presence of algorithms and tools such as Countdown, EAR (Energy Aware Runtime), and Examon, designed to address specific power management challenges, their limited interoperability and maintainability underscore the exigency for a comprehensive interoperability layer. These tools, though efficacious for specific use cases, often fail to cohesively integrate and cooperate due to varying interfaces and implementations.

Within this context, certain solutions have emerged, each with their own strengths and shortcomings. Some of the most well-known solutions include *Variorum* (LLNL), *GEOPM* (Intel), and *HDEEM* (Atos), to name a few. These solutions demonstrate efforts to address power management issues but exhibit varying degrees of interoperability. The need for a unified framework that bridges the gaps between different power management tools and techniques surfaces as a critical issue that necessitates exploration.

2.2 DDS

Drawing parallels from the field of robotics, ROS (Robot Operating System) confronted a similar dilemma and subsequently birthed a solution in the form of an interoperability layer built upon DDS. This successfully facilitated seamless communication among disparate robotic components.

In juxtaposing DDS's successes in robotics with the HPC domain, certain parallels emerge. Instances where a ROS-like approach thrives are evident, primarily in scenarios where uniform communication paradigms foster cooperation. However, challenges also manifest in areas where the robotics-inspired solution falls short of delivering adequate interoperability. The manifestation of these challenges serves as a motivation for more nuanced responses that cater to the unique intricacies of HPC power management.

Moreover, our exploration delves into the realm of DDS and its application within the power management context. The middleware implementation, exemplified by *rmw_dds_common* in ROS2, provides valuable insights into the practical implementation and utilization of DDS as a generalized middleware in real-world scenarios. This exploration becomes a stepping stone towards achieving the seamless communication and data coordination necessary for effective power management across distributed HPC environments.

3 Project's Description

The central theme of my PhD project will primarily analyze the performance of various DDS implementations, along with different configurations, in a quest to ascertain the most suitable approach. Once the optimal solution becomes apparent, the focus will shift towards leveraging of this implementation to enable interoperability among the different actors involved. An intriguing approach will be to employ the state-of-the-art framework proposed by ROS2, along with its Ros-Middleware (*rmw*).

A second pivotal aspect of the project will encompass the actual implementation of the previously developed middleware across all components constituting the HPC powerstack. This implementation will shed light on an **"interoperable runtime for distributed power management of large-scale HPC systems based on DDS"**, facilitating the seamless management of power on a significant scale within a distributed environment. This endeavor aims to solidify the interconnection between various actors, enabling dynamic and collaborative resource management within the realm of high-performance systems.

A middle step will be the development of a empiric simulation, needed to test and compare different level of leverages (bare DDS with DDS middleware) considering also increasing entities involved in the communication process.

4 Expected Results

This project is expected to achieve three main contributions:

- the analysis of several advanced control structures aimed at answering the new upcoming requirements and constraints originated from the increasing complexity of processors in the edge and high-performance computing domains;
- The exploration of more compute-intensive control paradigm (model predictive control, deep reinforcement learning), and HW architectures to execute them efficiently.
- the development of a simulation framework for the co-design of all the parts of the control system (HW, run-time/RTOS, control policy, interfaces with processors and sensors);
- the development of an open-source power controller firmware based on open-source hardware, aimed at being implemented in a broad range of chips.

5 Proposed project timeline

- Year 1:
  - Literature overview on advanced control designs and on the State of the Art control algorithms applicable to the power and thermal control of a processor.
  - Creation of a simulation framework. Development of system and controller models.
  - Initial Firmware architecture design.
- Year 2:
  - Analysis and comparison of the identified control structure designs.
  - Intermediate version of the controller Firmware targeting an identified open-source hardware design.
- Year 3:
  - SECURITY?
  - Adapt the project to a broad range of scenarios and possible implementations.

6 Outline of the proposed findings assessment criteria

The criteria to asses the proposed findings will be:

- an open-source publication on GitHub of an operational firmware for a processor power controller, associated with an open-source RTL component regarding the controller hardware.
- the possibility to consider the implementation of the open-source controller project in a real processor design.
- an exhaustive analysis of the control structures relevant to the power and thermal control of a processor, publishable on a paper.
- the development of a simulation framework able to provide reliable results.