July 2017

Energy Saving and Thermal Management Opportunities in a Workload-Aware MPI Runtime for a Scientific HPC Computing Node

Abstract. With the advent of a new generation of supercomputers characterized by tightly-coupled integration of a large-number of powerful processing cores in the same die, energy and temperature walls are looming threats to the growth in computational power.

Scientific computing is characterized by a single application running in parallel on multiple nodes and cores until termination. The message-passing programming model is a widely adopted paradigm for explicitly handling data-sharing between processes of the same application. As an effect of the MPI communication patterns among different processes, the application is characterized by phases which can be exploited by OS power manager. In addition, the large number of cores integrated in the same silicon die introduces large thermal capacitance as well as on-die thermal heterogeneity. Jointly exploiting local workload unbalance and computational node heterogeneity can open interesting opportunities for advanced thermal and energy management. In this paper, we present an exploratory work to assess these opportunities and their limiting factors. We analyze application workload and we identify opportunities to reduce energy consumption and their impact on performance. We test our methodology on a widely-used quantum-chemistry application demonstrating potential benefits of combining the application flow with power and thermal management strategies.

Keywords. HPC, thermal model, power model, energy, MPI, runtime, scientific workload

1. Introduction

Nowadays, it is well established that the pace dictated by the Moore's law on technological scaling comes at the cost of increasing power consumption and leads to thermally-bound computing systems. Supercomputers as well as data centers are on the cutting edge of this crisis, because of aggressive performance, integration density and sustainable power budget [16,19].

The most powerful supercomputer in Top500 is *Sunway TaihuLight* which consumes 15.3 MW to deliver 93 PetaFLOPs. The second one, *Tianhe-2* (ex 1st) consumes 17.8 MW for "only" 33.2 PetaFLOPs. However, the power consumption increases to 24 MW when considering also the cooling infrastructure [9]. Such an amount of cooling power serves to prevent thermal issues. Increasing the inlet coolant temperature reduces the cooling cost, but impacts thermal budget. Druzhinin et al. [17] reports a performance drop of 10% caused by thermal throttling when raising inlet cooling water from 19^oC up to 65^oC in a direct liquid cooling supercomputer.

Beneventi et al. [4] show in an Intel-base computing node with 36 physical cores, that the increased number of processors integrated on same die generates significant thermal gradients and heterogeneity. On the same silicon die, they measure up to 24^oC of temperature difference between active cores and idle cores, and more than 7^oC of thermal heterogeneity under homogeneous workload.