2A-03

Upgrading Job Scheduling by Disaggregating Rack Resources in Datacenters

Yao Hu[1], Ikki Fujiwara[2], Michihiro Koibuchi[1]

[1]National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan 101-8430
{huyao, koibuchi}@nii.ac.jp
[2]National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Kyoto, Japan 619-0289
ikki@nict.go.jp

Abstract

This study provides a comparative evaluation of job scheduling on two resource disaggregated datacenter infrastructures, i.e., a rackscale (RS) infrastructure and an inter-rackscale (IRS) infrastructure. The former emerges as a famous custom architecture where one rack consists of a number of processors, storages and accelerators that can be customized to a target application. The latter is further designed as an evolution of the former to disaggregate various hardware resources into different racks according to their own areas. For RS, its resource utilization is low because a whole rack is exclusively occupied by one job. For IRS, it uses free-space optics (FSO) for tightly-coupled connections between processors, storages and GPUs distributed in different racks, by swapping endpoints of FSO links to change network topologies. Through our evaluation by a large system simulation, we present the advantage of the FSO-equipped IRS architecture in job scheduling such as average queuing time of all dispatched jobs.

Keywords: Rackscale architecture, interconnection network, free-space optics, datacenter, job scheduling

1. Introduction

Current datacenter servers are not optimally configured for each specific application, which can result in resource waste and inefficiency. Hence, traditional datacenters are under severe pressure to meet growing demands of cloud, big data, mobile and social collaboration applications.

To deal with the system underutilization problem due to resource aggregation in one rack server, two resource disaggregated datacenter infrastructures are proposed (Figure 1). One is rackscale (RS) [1] [2] proposed recently which disaggregates compute, storage and network resources and introduces the ability to pool these resources for more efficient utilization in a machine room. It simplifies resource management and provides the ability to dynamically compose resources based on workload-specific demands. The RS computer systems can be customized to various types of newly generated applications.

Another competitive datacenter architecture is inter-rackscale (IRS), which is designed to allocate hardware resources to serve different types of applications simultaneously so as to improve overall system utilization. IRS disaggregates various hardware resources into different racks according to their own areas. The resources of one IRS rack can be shared by multiple different user applications.

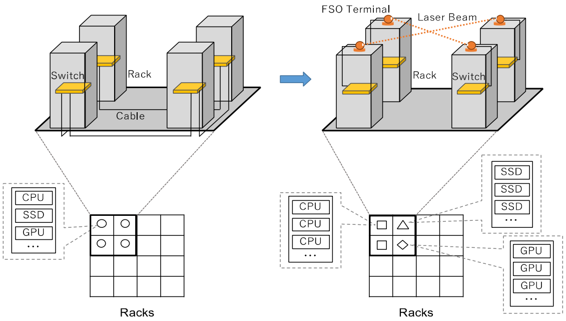

Figure 1. The RS architecture and the IRS architecture.

The heart of IRS is to use a wireless technology called free-space optics (FSO) [3] with high bandwidth over 10 Gbps for communication between racks. The wireless IRS system assume that the inter-rack links are wireless while the intra-rack links are electric or optical cables. An important property of the wireless system is that the link endpoints can be swapped on demand so as to lower communication latency, which makes job mapping and scheduling more flexible. Since the job mapping of various parallel applications becomes important as the scale of high-performance systems increases, an effective interconnection network of IRS racks can potentiate resource utilization and application optimization.

In this study, we attempt to provide a comparative evaluation of job scheduling on the RS architecture and the IRS architecture. The evaluation results show that the job scheduling performance such as average queuing time can be upgraded by disaggregating rack resources under the IRS architecture.

2. Methodology

We consider the job mapping and scheduling over 3-D Torus interconnection networks. According to geographical rack floor-layouts we compare two prototypes for IRS over 3-D Torus, namely IRS-REPEAT and IRS-LOOP (Figure 2).

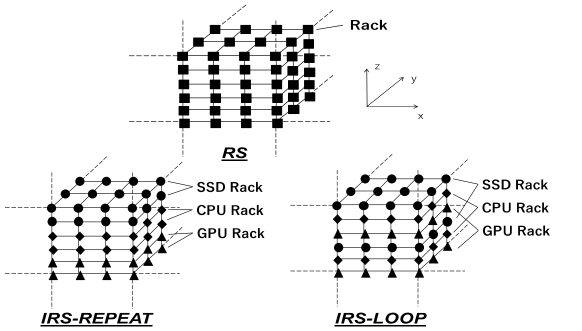

Figure 2. The RS racks and IRS racks over 3-D Torus interconnection networks.

We assume a job mapping on a specific topology for the inter-CPU network and on a Star for the inter-CPU/SSD(GPU) network. First, each job is assigned to a set of unused compute racks that form a specific network. Then, CPUs require to transfer data with SSDs and/or GPUs for job execution. The arrangement of racks with diverse architectures included in our evaluation is shown in Table 1.

Table 1. The arrangement of racks with diverse architectures over 3-D Torus interconnection networks.

Architecture	Racks	Arrangement of Racks	FSO(s) per Rack
RS	3072	16×16×12	-
IRS-REPEAT	3072	16×16×12 (repeat on z axis)	0/1/2/3/4
IRS-LOOP	3072	16×16×12 (loop on z axis)	0/1/2/3/4/5/6

3. Evaluation

We evaluate the performance of a large RS/IRS system with NPB applications. To acquire execution times (simulation cycles) for the NPB applications running on diverse guest topologies (2-D Mesh, 3-D Mesh, 3-D Torus and Random), we first get a series of simulation results by SimGrid with different number of processors (4, 16, 64 and 256). We assume that the host network topology is 3-D Torus (16×16×12).

We generate n = 2000 jobs as workloads with random arrival timings for a Poisson process with λ = n/1000. Each job is given priority to be mapped to the system with dilation-1. The dilation-2/3/4 mapping is also considered if the current dilation-1 mapping cannot be found on the system. We also take into account communication latency overhead of inter-CPU/SSD(GPU) data transfers.

Figure 3 shows the simulation results of the average queuing time of all dispatched jobs. It can be seen that the IRS architecture (either IRS-REPEAT or IRS-LOOP) outperforms the RS architecture because IRS offers a finer resource granularity for job allocation due to its entirely disaggregated server environment.

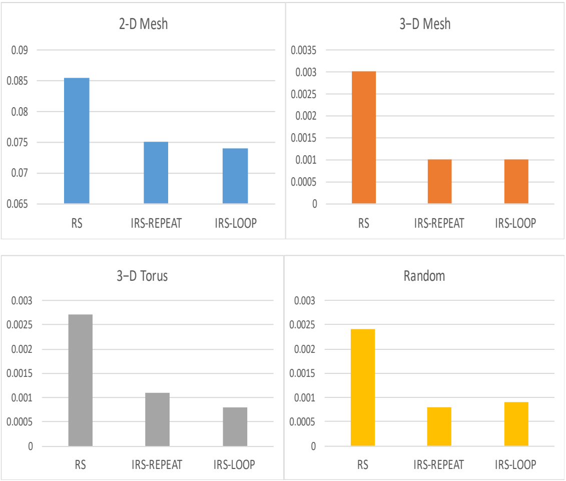

Figure 3. The average queuing time of all dispatched jobs.

4. Conclusion

This study provides a comparative evaluation of HPC job scheduling on the rackscale (RS) infrastructure and the inter-rackscale (IRS) infrastructure, which offers a finer resource granularity for job allocation due to its entirely disaggregated server environment. The simulation results show that IRS outperforms RS in job scheduling such as average queuing time of all dispatched jobs.

Acknowledgement

This work was partially supported by MIC/SCOPE (152103004) and KAKENHI #16H02816.

References

[1] Intel Rack Scale Architecture: Faster Service Delivery and Lower TCO. [Online]. Available: http://www.intel.com/content/www/us/en/architecture-and-technology/intel-rack-scale-architecture.html.

[2] S. Legtchenko, N. Chen, D. Cletheroe, A. Rowstron, H. Williams, and X. Zhao, "Xfabric: A Reconfigurable In-rack Network for Rack-scale Computers," in NSDI'16 Proceedings of the 13th Usenix Conference on Networked Systems Design and Implementation, 2016, pp. 15–29.

[3] I. Fujiwara, M. Koibuchi, T. Ozaki, H. Matsutani, and H. Casanova, "Augmenting Low-latency HPC Network with Free-space Optical Links," in 21st International Conference on High-Performance Computer Architecture (HPCA), Feb. 2015, pp. 390–401.