

A Reliability-Aware Resource Allocation Method in Disaggregated Data Centers

 Communications & Information

Computer/AI/Data Processing and Information Technology
 Digital Broadcasting, Telecommunication and Optoelectronics

中文版本

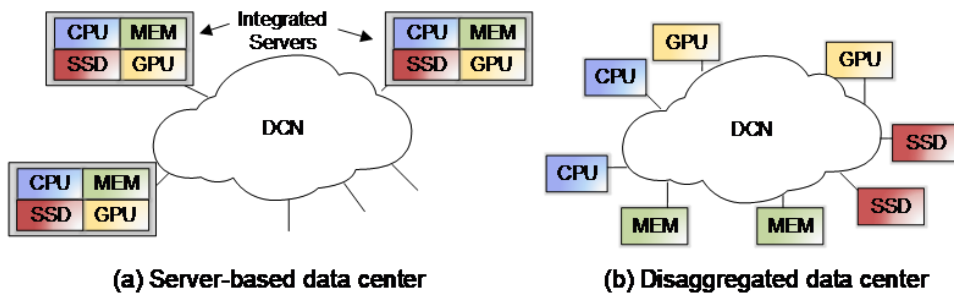


Fig. 1. Server-based and disaggregated DCs.

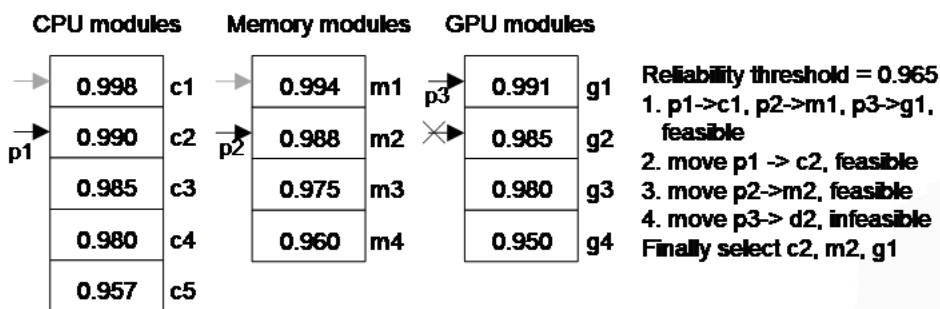



Fig. 2. An illustrative diagram of exemplary settings and conditions of an execution of service request to a DDC implemented with a heuristic process in accordance to one embodiment of the present invention.

Opportunity

Data center (DCs) are playing an ever more important role in providing Internet services. The global data center (DC) market was valued at USD44.42 billion in 2020 and is expected to grow by 13.3% from 2021 to 2028. A large-scale DC consists of numerous servers, each of which tightly integrates various resources. These servers are interconnected by a dedicated network.

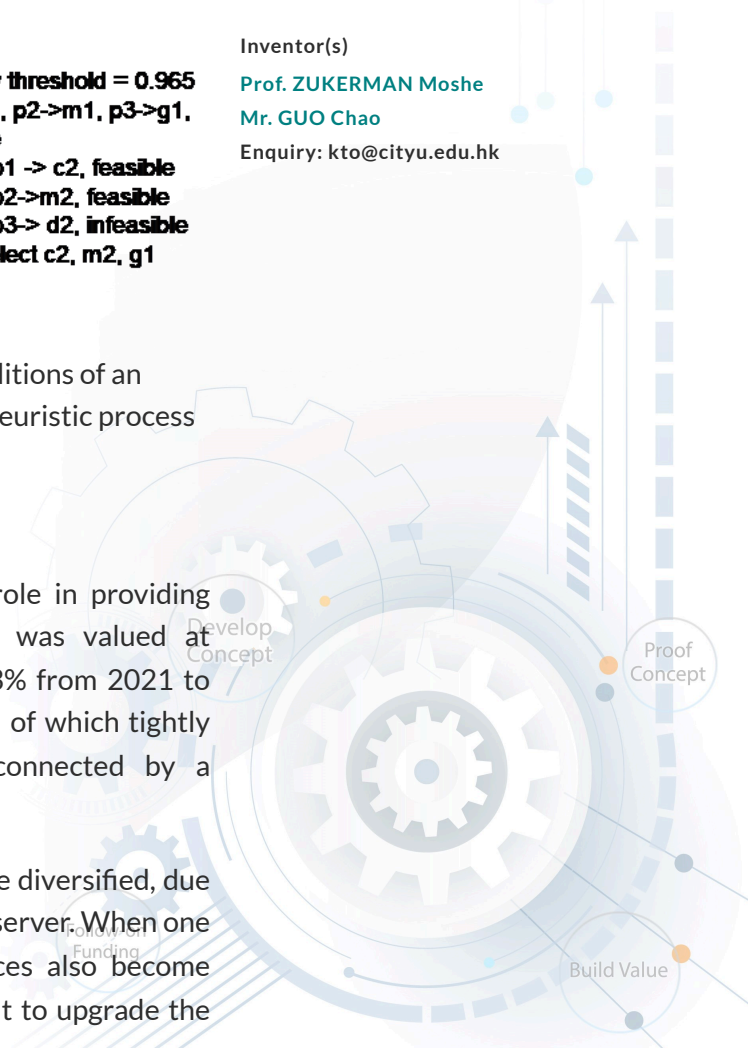
However, this architecture is inefficient when workloads are diversified, due to the close coupling of different types of resources in each server. When one type of resource in a server is exhausted, other resources also become unavailable and are then wasted. This also makes it difficult to upgrade the

IP Status
 Patent filed



Technology Readiness Level (TRL) ? 4

Inventor(s)
 Prof. ZUKERMAN Moshe
 Mr. GUO Chao
 Enquiry: kto@cityu.edu.hk



server. Although components in a server often have very different lifecycles, hardware upgrades and resource expansions are typically done only at the server level.

With the rapid evolution of Internet services and user demand, hardware and software are changing rapidly, requiring more and more frequent hardware refreshing. To reduce refreshing costs and improve resource utilization and upgradability, resource disaggregation is an increasingly prominent solution. Through hardware decoupling, different types of resources can be upgraded or scaled independently and agilely, which significantly reduces costs. Novel methods of resource allocation are urgently needed to support this emerging technology.

Technology

This invention presents a pioneering method of reliability-aware resource allocation for the data centers (DC) of the future. First, the researchers modelled the reliability of a resource allocation request in a server-based or disaggregated DC. They then considered a resource allocation problem to maximize the number of requests accepted with guaranteed reliability. This was formulated as an integer linear programming problem. In addition, they proposed a more straightforward heuristic approach. They conducted a large number of simulation studies, and their numerical results revealed that it may be possible to significantly improve the service reliability of DCs using their novel resource-disaggregation approach.

Advantages

- Better guarantees service reliability in DCs
- Greater resource efficiency in DCs

Applications

- Use by cloud computing providers investing in building large-scale DCs
- Realizing reliability-guaranteed resource allocation while maximizing profits or minimizing costs for DC owners, e.g., Microsoft Azure and Amazon Web Services.

