Memory Management in the Cloud

Saurabh Mathur
saurabhmathur96@gmail.com

SITE, VIT Vellore

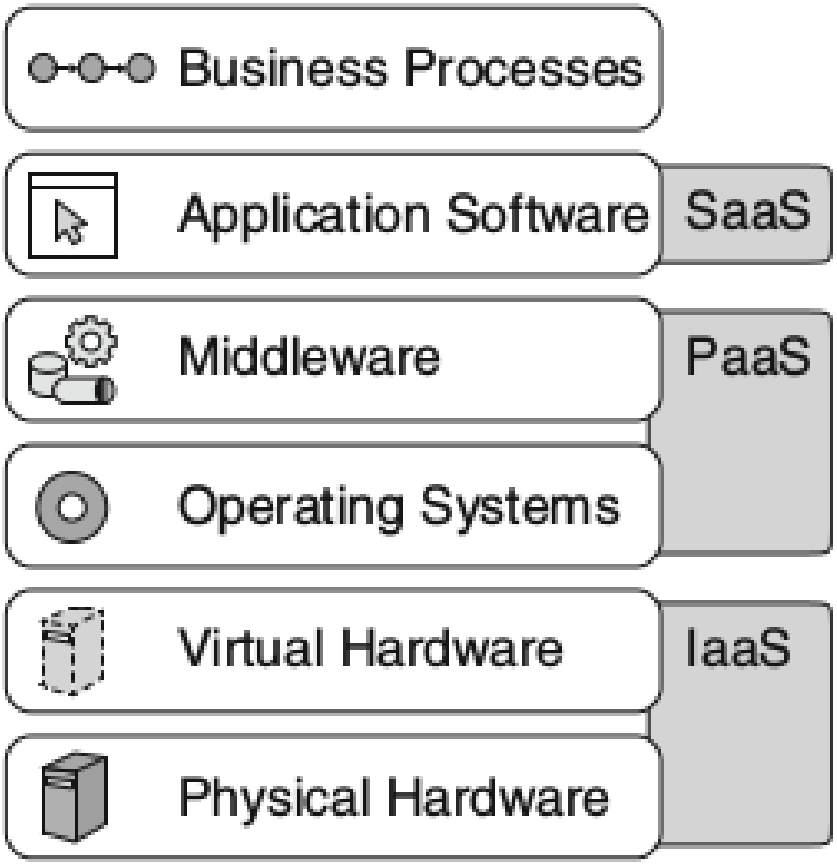
April 28, 2016

Definition

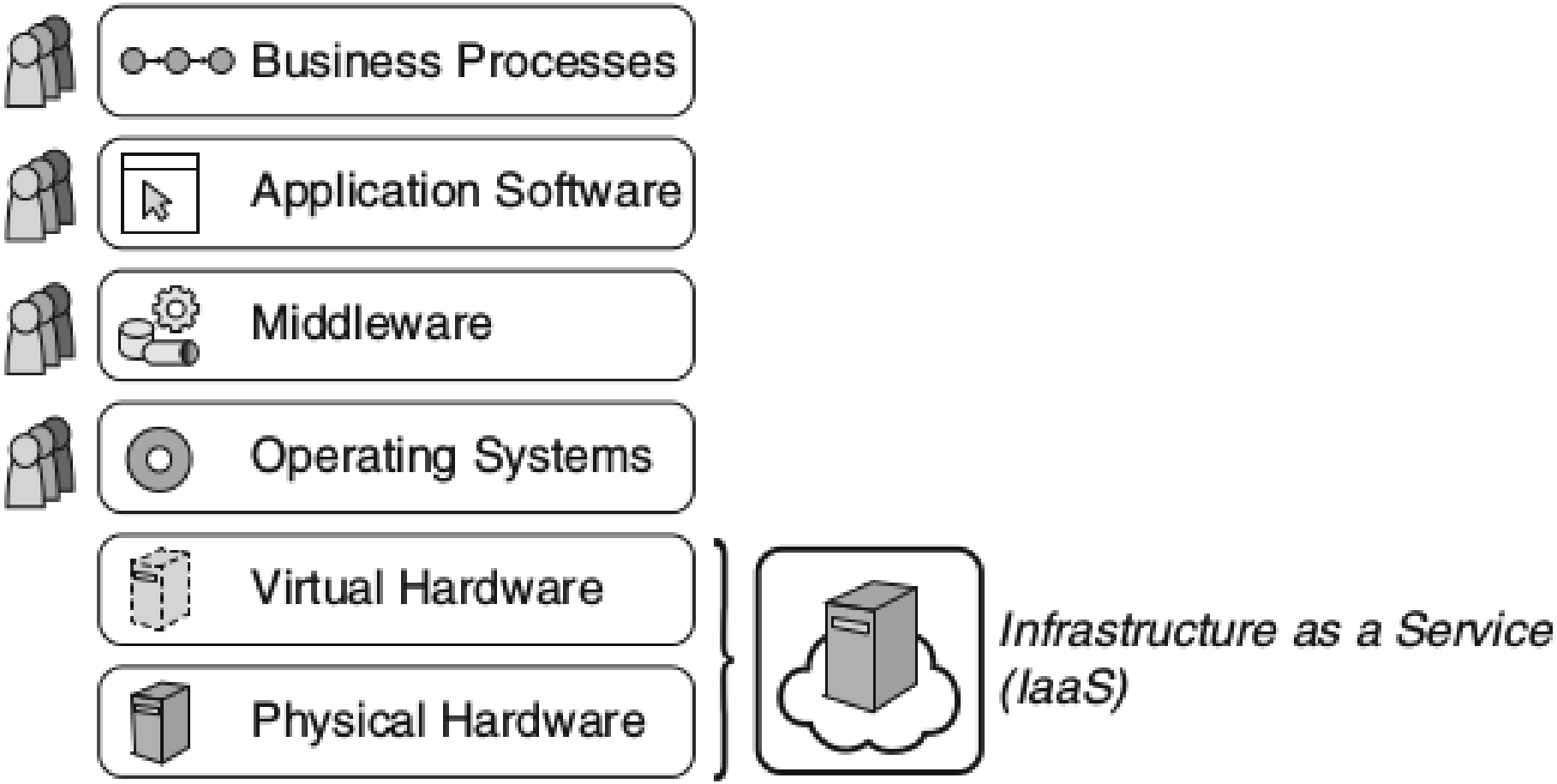
Cloud Computing

A kind of Internet-based computing that provides shared processing resources and data to computers and other devices on demand.

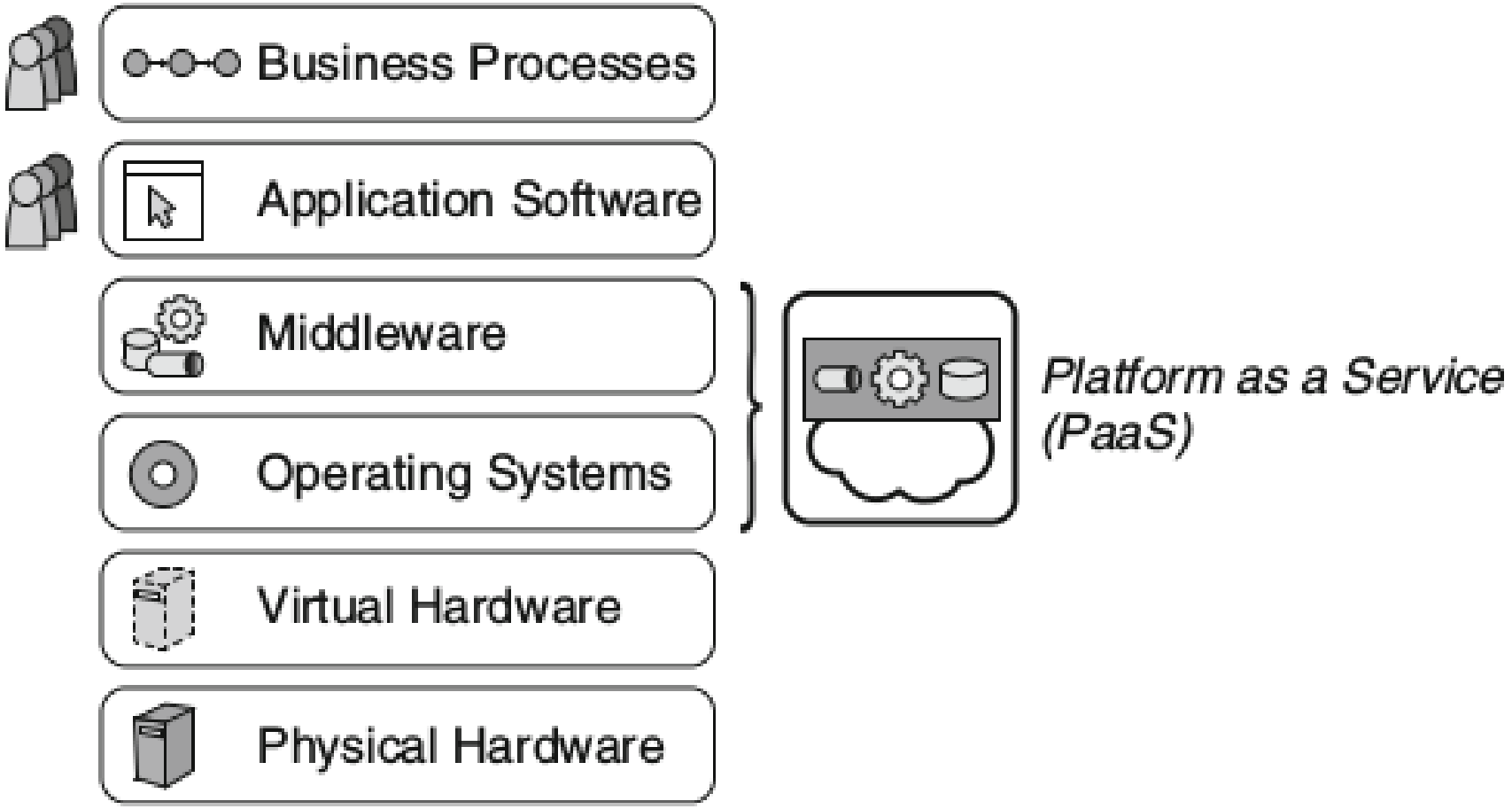
The Cloud Computing Stack



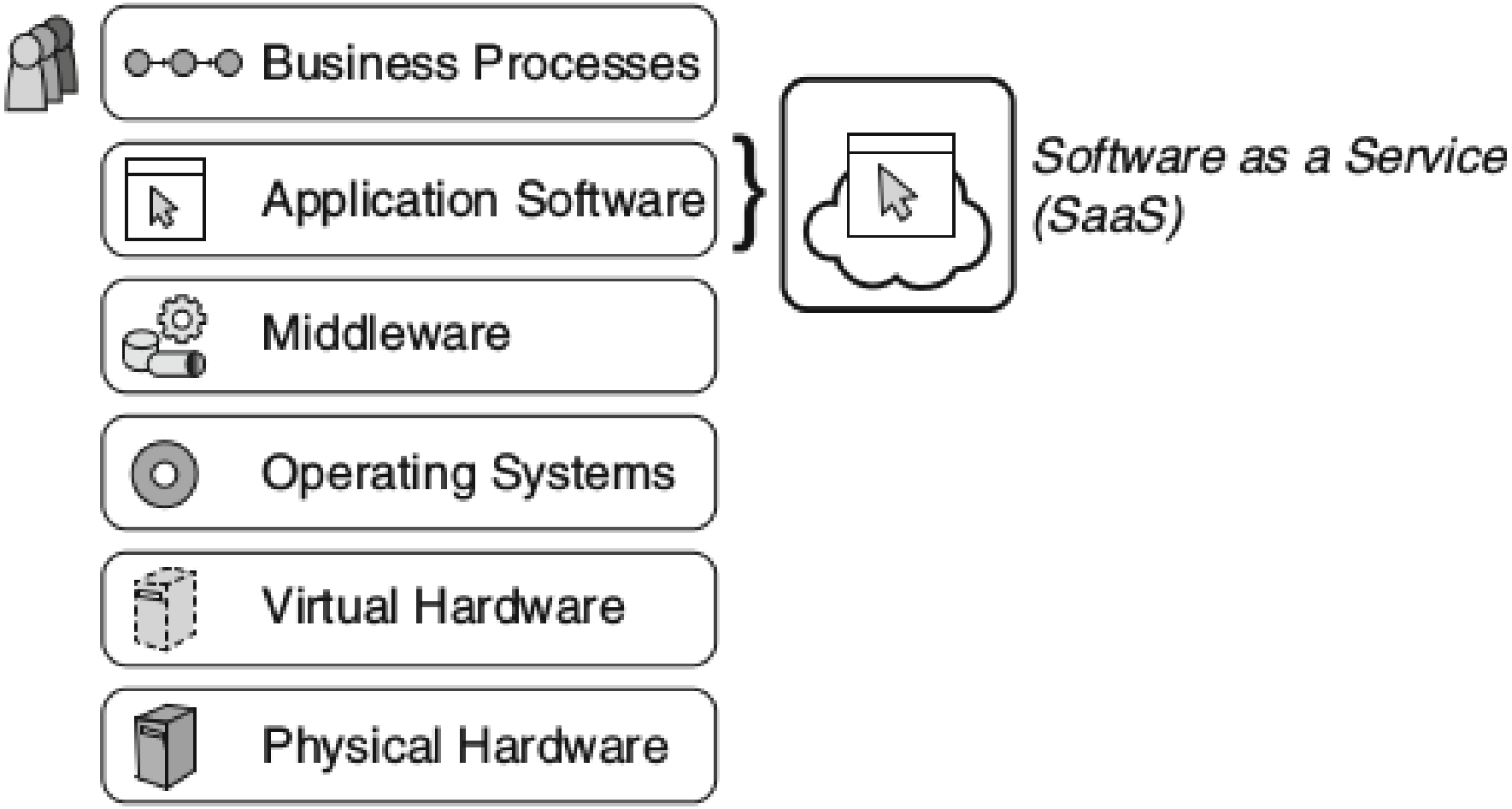
The Cloud Computing Stack



The Cloud Computing Stack



The Cloud Computing Stack



Characteristics

- Reduction in capital expenditure
- Device and location independence
- Sharing of resources (and costs)
- Centralization of infrastructure and data
- Reliability by way of multiple redundant sites
- Scalability

Memory Management

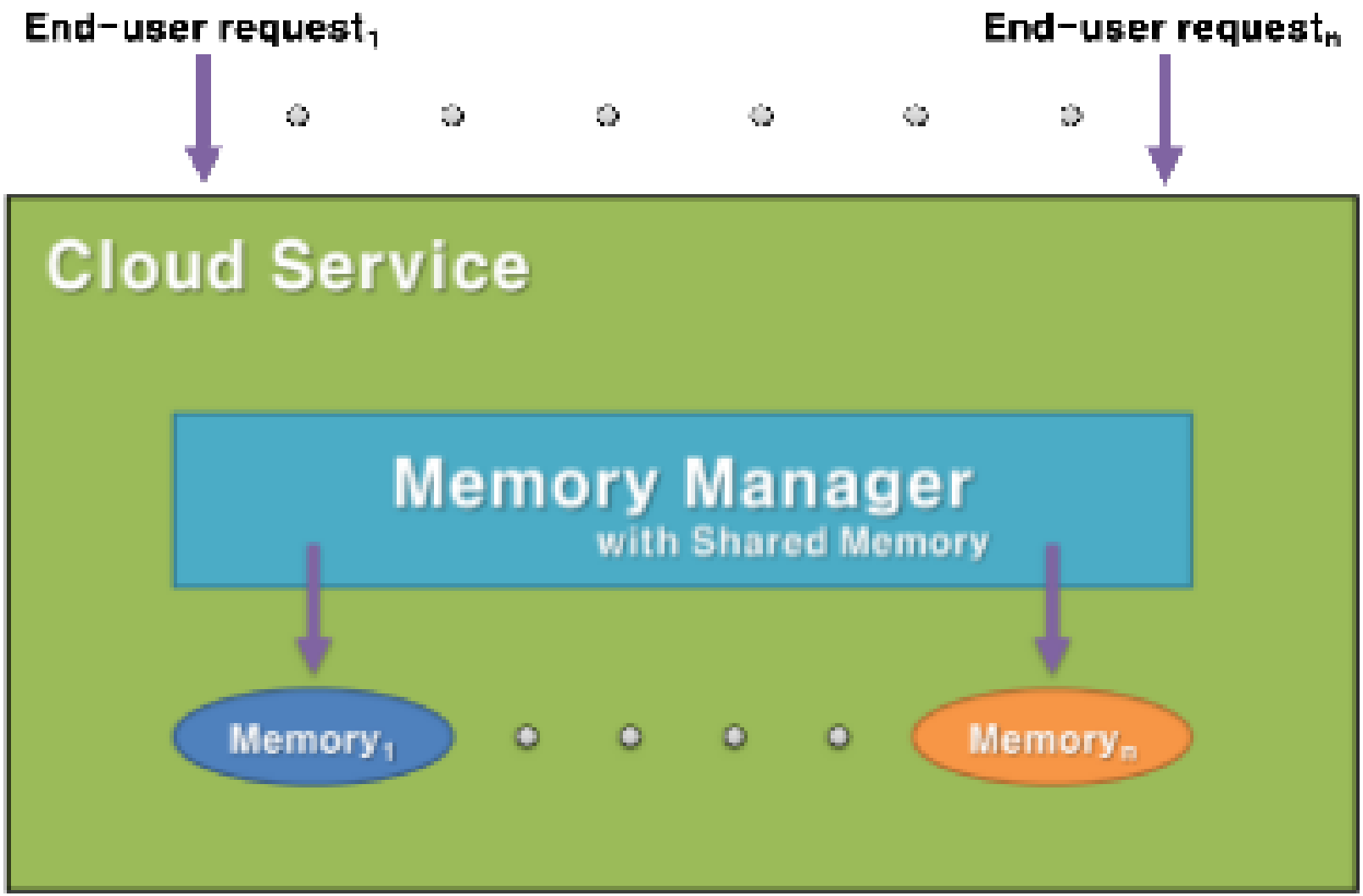


Fig. 1. Target Architecture Overview

Page Replacement

Fig. 2. Conventional Management Policy

Static Partitioning

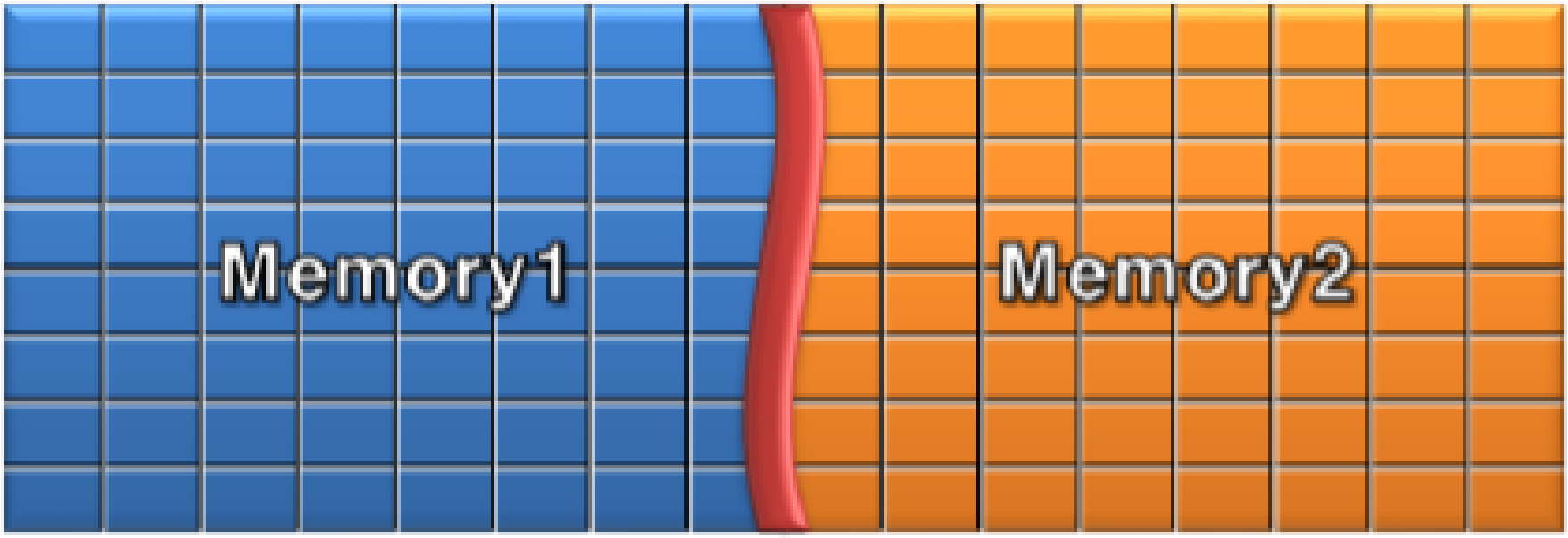


Fig. 3. Static Partition Management Policy

Static Partitioning

- Every memory is divided statically by predefined partition ratio
- Simple implementation
- Less Overhead
- Calculation needs to be done beforehand
- Memory is wasted

Dynamic Partitioning

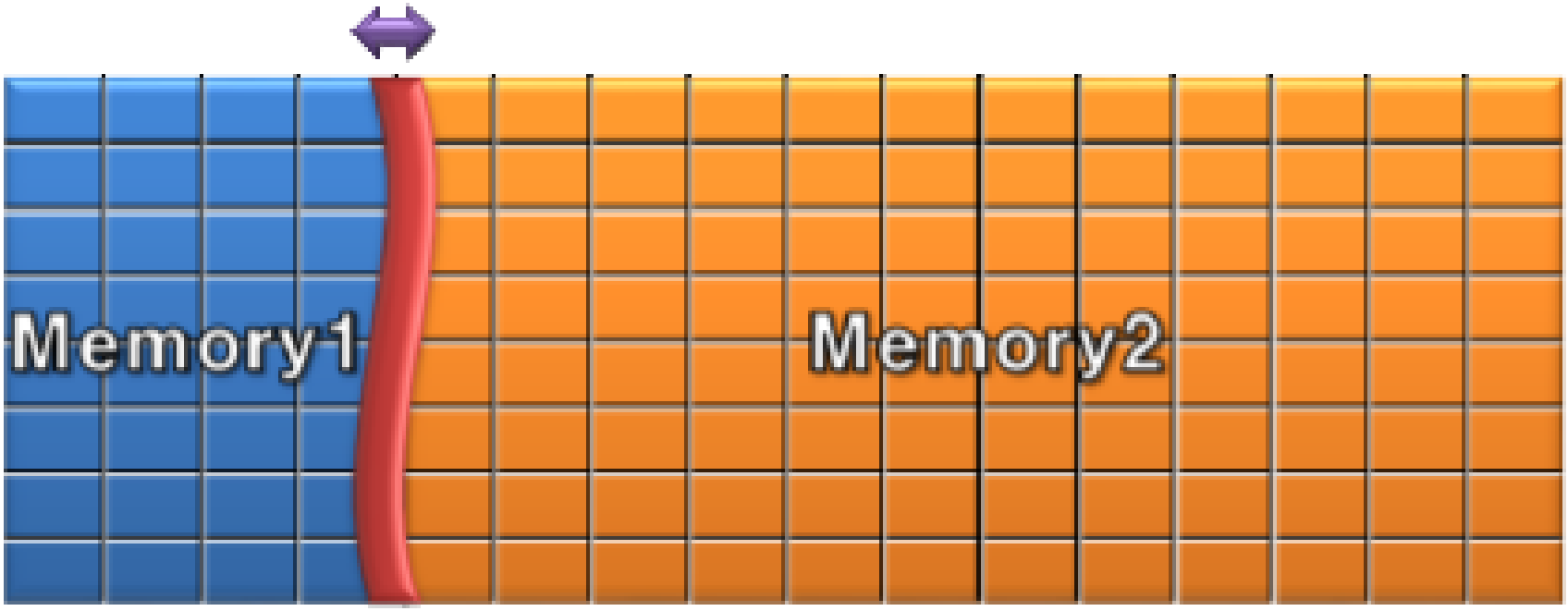


Fig. 4. Dynamic Partition Management Policy

Dynamic Partitioning

- Every line of shared memory is divided dynamically depending on the number of misses for each memory
- Efficient in terms of memory
- Overhead due to calculation
- No calculation required before deployment
- 4.65% performance boost

Static Partitioning - Statistics

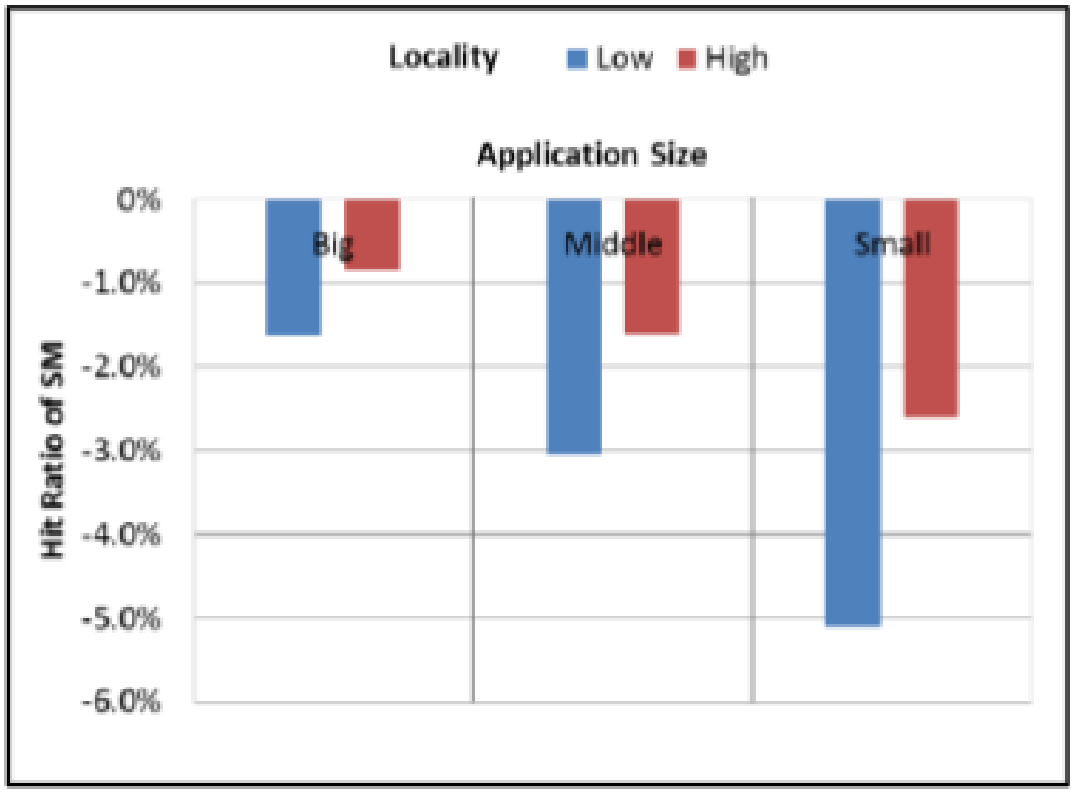


Fig. 5. Hit ratio for Static Partition Management Policy

Dynamic Partitioning - Statistics

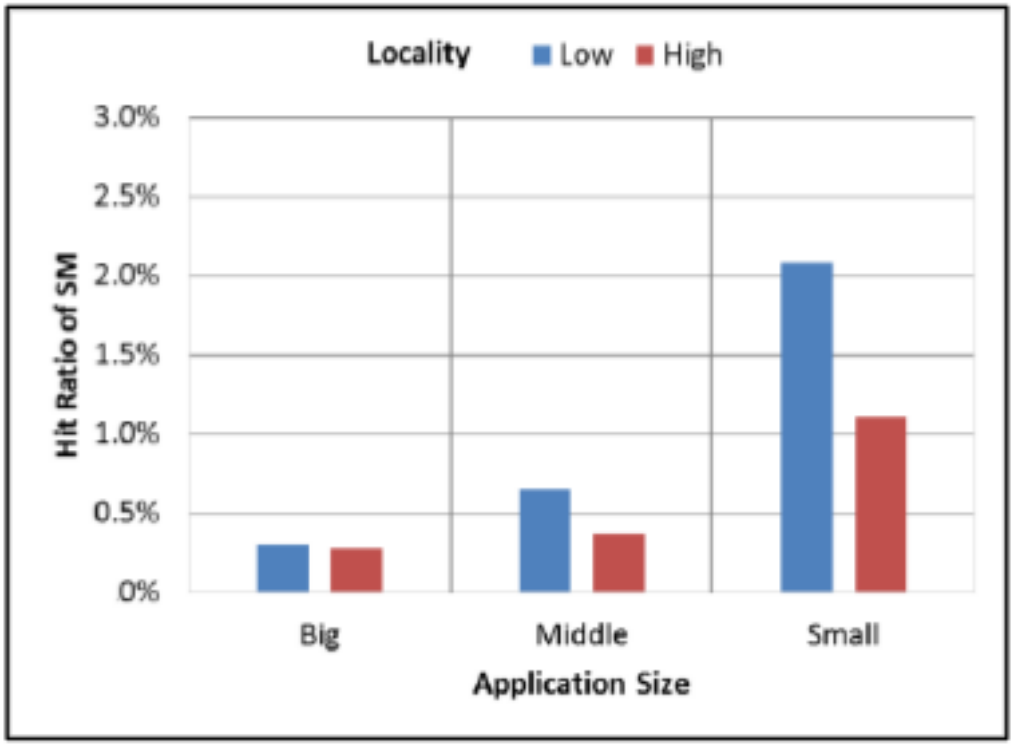


Fig. 6. Hit ratio for Dynamic Partition Management Policy

References

- *Analysis of Memory Management Policies for Heterogeneous Cloud Computing*, Dong Oh Son et al.
- *Cloud Computing Patterns*, C. Fehling et al.