Fair & Elastic Resource Allocation in Cloud Computing Environments

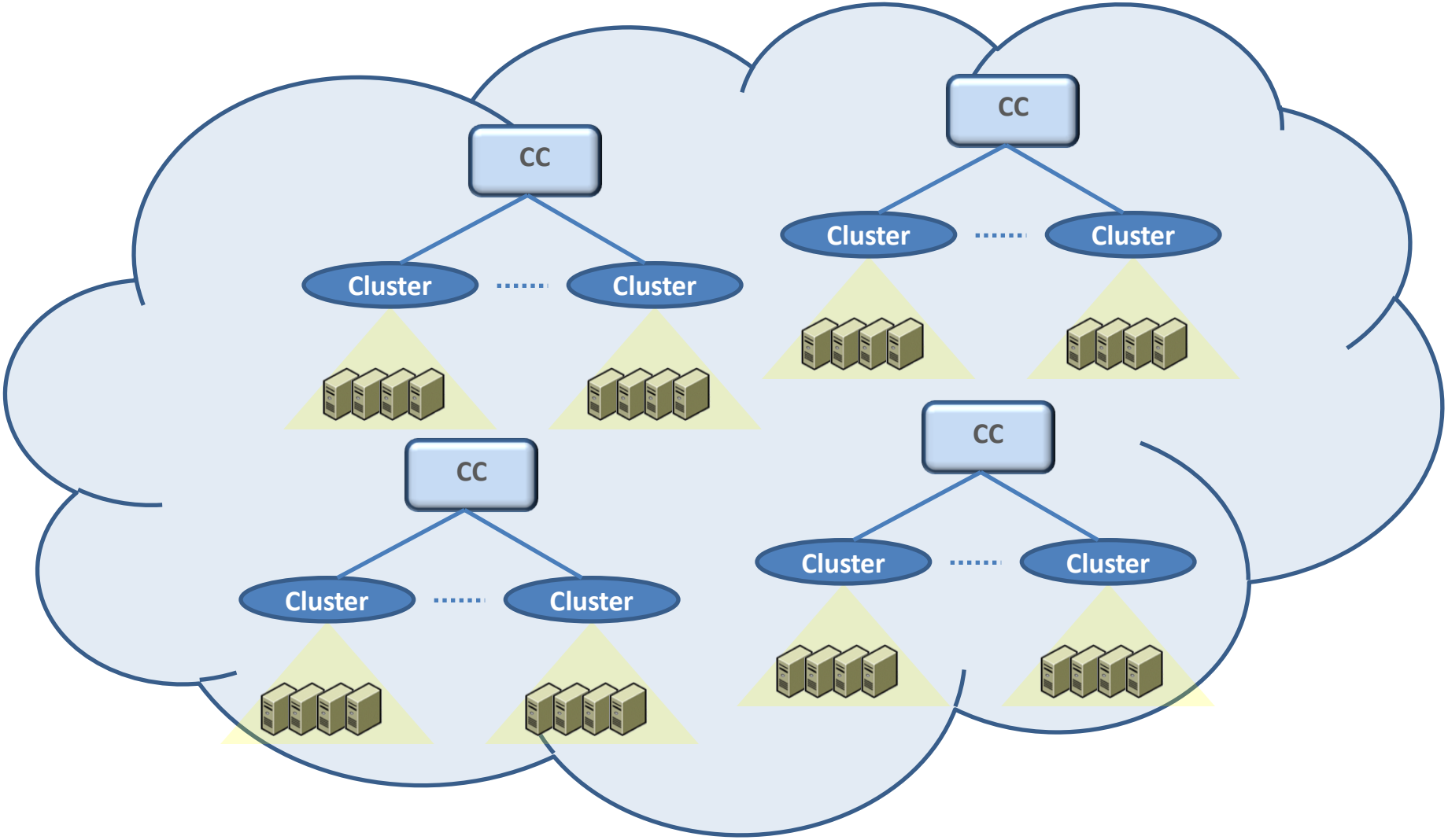
GT: **Mukil Kesavan**, Ada Gavrilovska, Karsten Schwan

VMware: Orran Krieger, Irfan Ahmad, Ravi Soundararajan

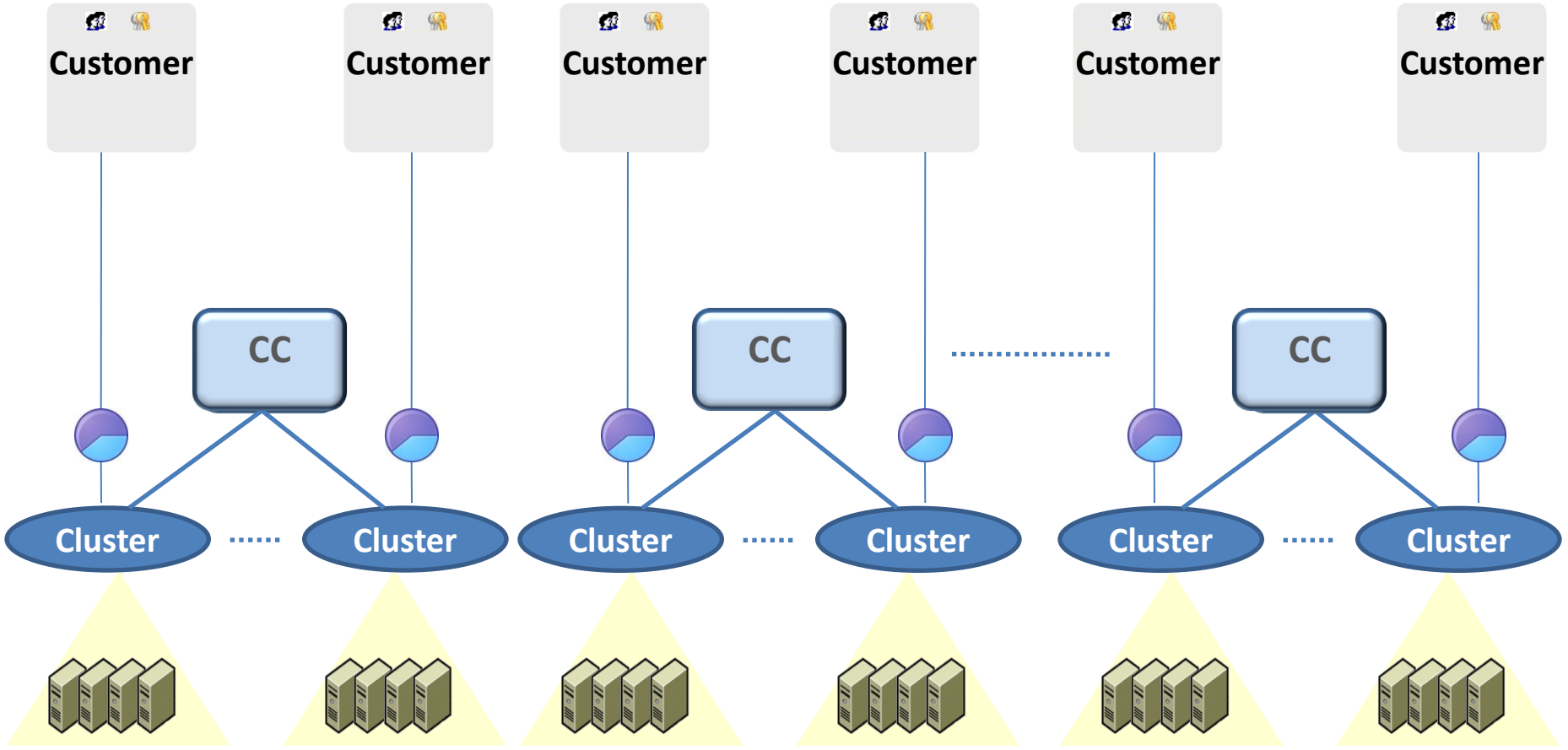
Goals

- Scalable resource management ~ 10k hosts
- Load-Balance Resource Consumption
- Work for Broad Class of Workloads
- Elasticity: Demand based allocation of resources

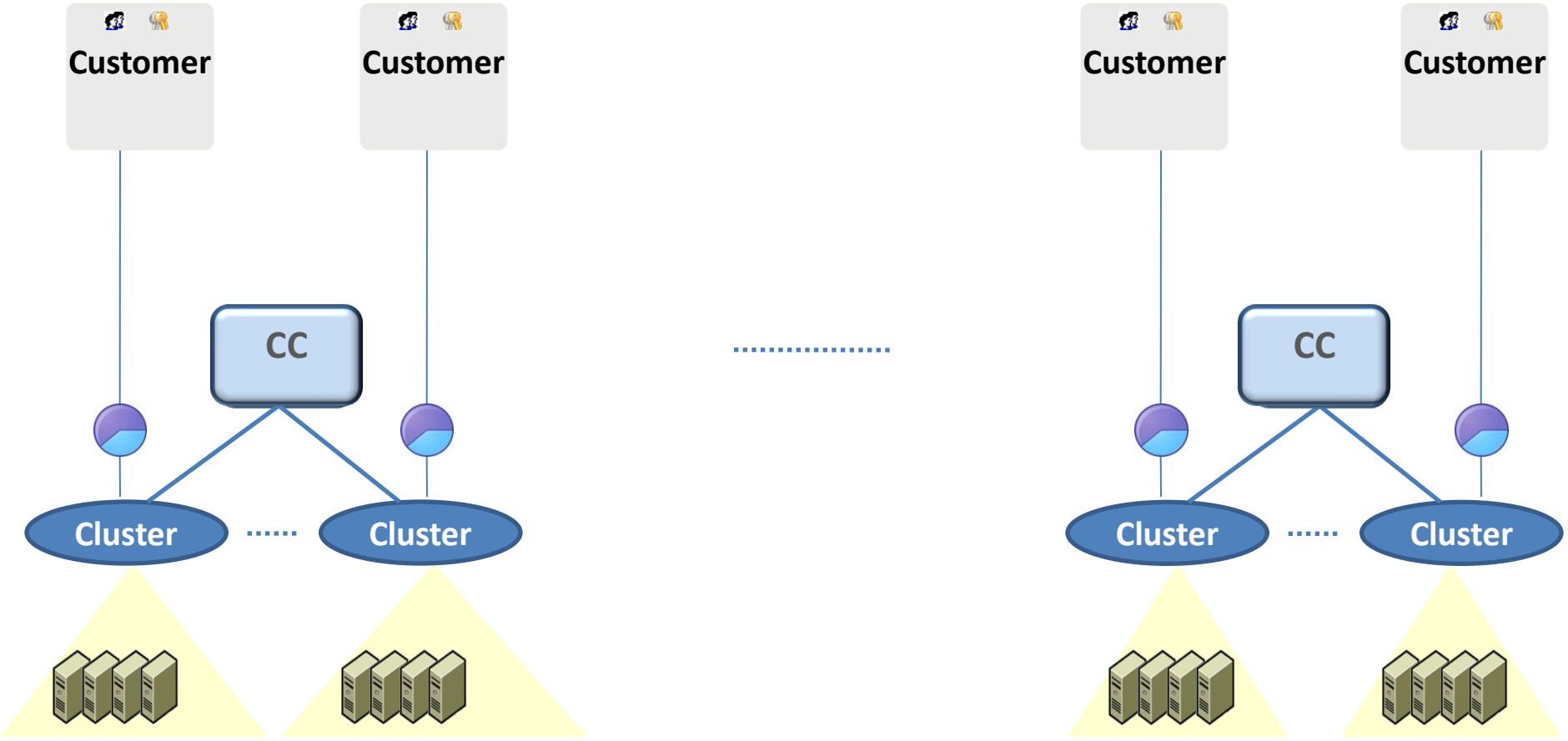
Cloud Architecture



Imbalance across Cloud



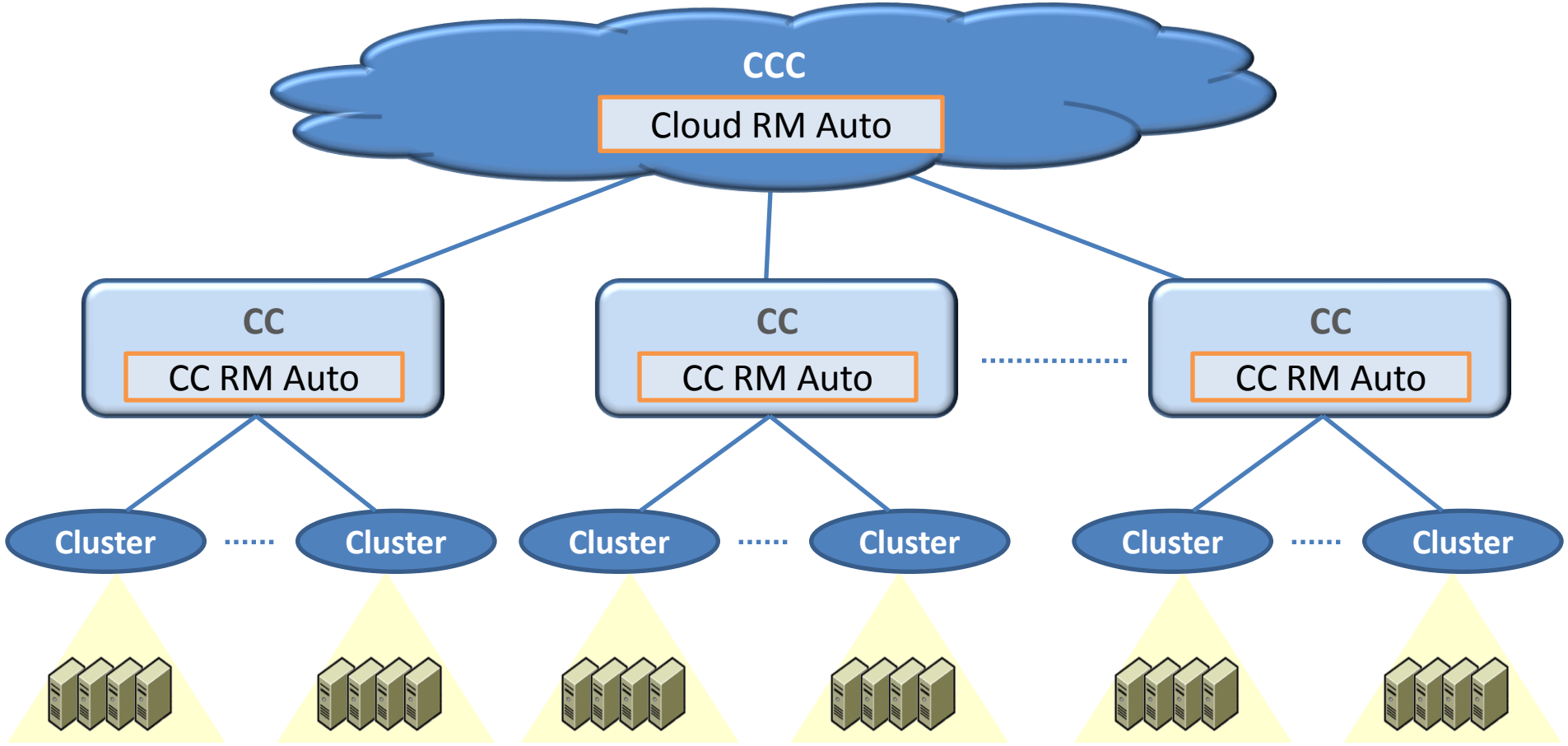
Dealing with Imbalance



Automation Actions

- Move capacity across management hierarchy
  - Preserve association of VMs to management agents
  - Granularity: Hosts
- Deploy existing centralized solutions with management agents

Hierarchical RM Architecture

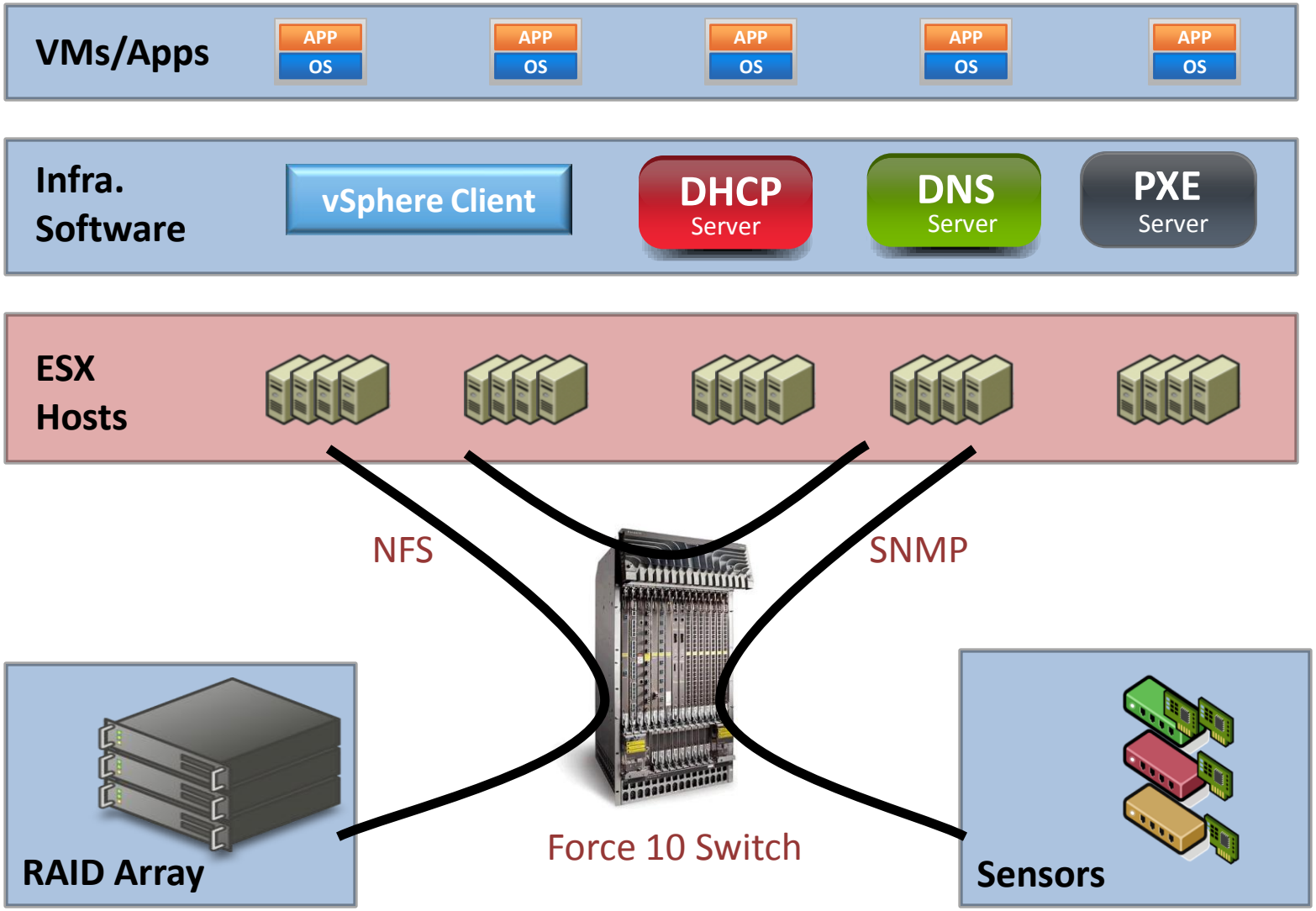


RM scalability achieved by two additional levels of automated load balancing

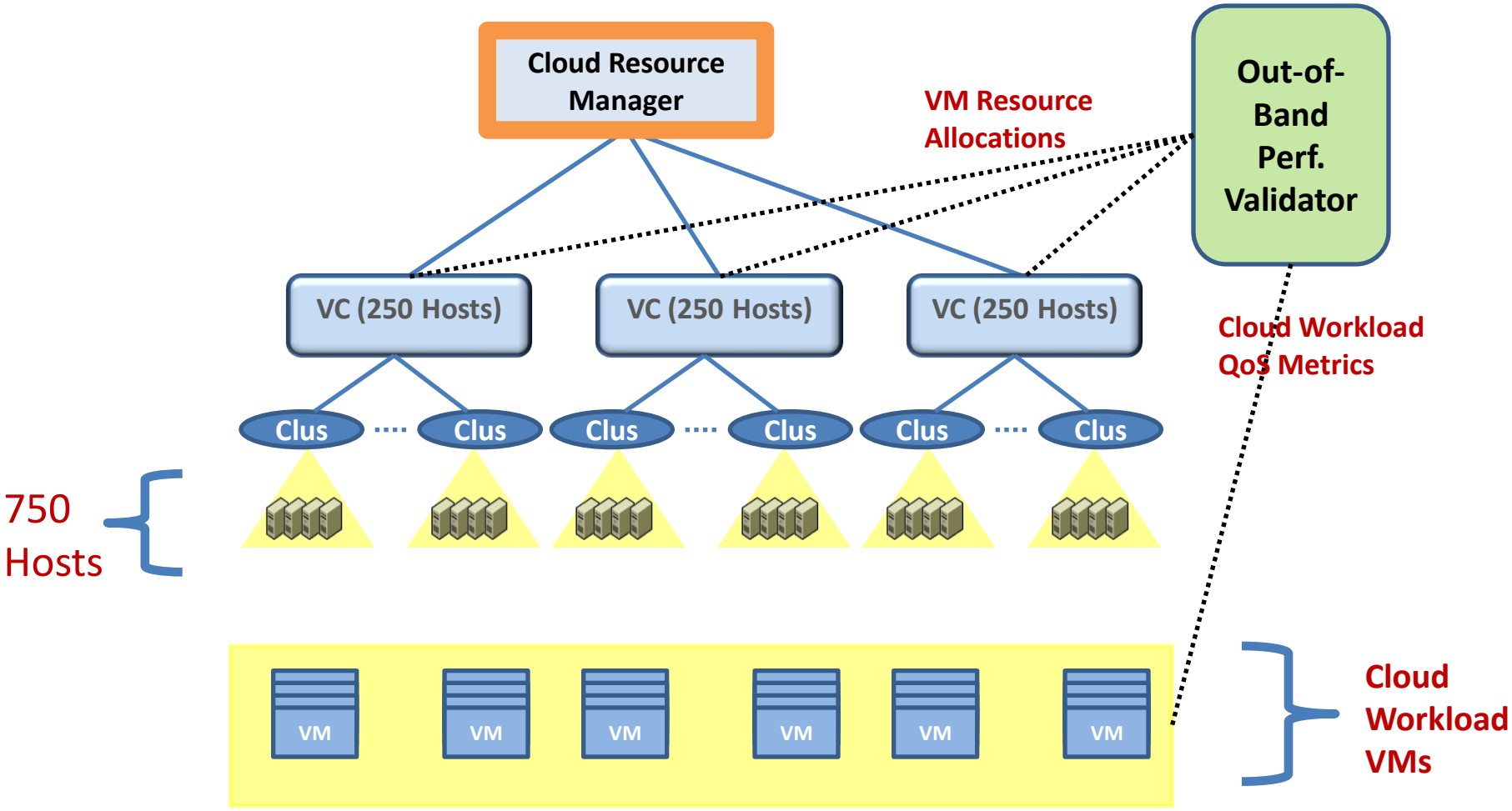
Algorithm Design

- Resource Demand based allocation
  - Measure, Aggregate & Predict
- Honor Static Constraints
  - Reservations, Limits, Fault Tolerance etc.
- Time Scales of Operation
- Scalability
- Host Selection

Techway Infrastructure Setup



Load Balancing Testbed



Benchmarks

1. *VMMark*
2. *Hadoop*
3. *Nutch*: Map-Reduce Web Search
4. *Voldemart/YCSB*: Distributed K/V Store
5. Linpack HPL: MPI Based HPC Code
6. *Berkeley CloudStone*: Self-Scaling Web

Evaluation Plan/Metrics

1. Cloud Resource Utilization
   - Aggregate Utilization of Powered On M/Cs
2. Load Imbalance across CCs & Clusters
   - Steady State Convergence
3. Adaptation to Changes/Spikes in Res. Usage
4. Workload Quality Metric (e.g. Search Query Time)
5. Algorithm Overhead – Enforcing Mgt. Actions

Other Research Projects

- Storage I/O Allocation with Isolation at the App Level
- Black-box Monitoring & VM Ensemble Detection
- Power-centric Mgmt & Billing
- Generic, Flexible Management Overlays

Thank You