Cost Reduction in Hybrid Clouds for Enterprise Computing♪

Biyu Zhou, Fa Zhang, Zhiyong Liu (ICT, CAS)
Jie Wu (Temple University)

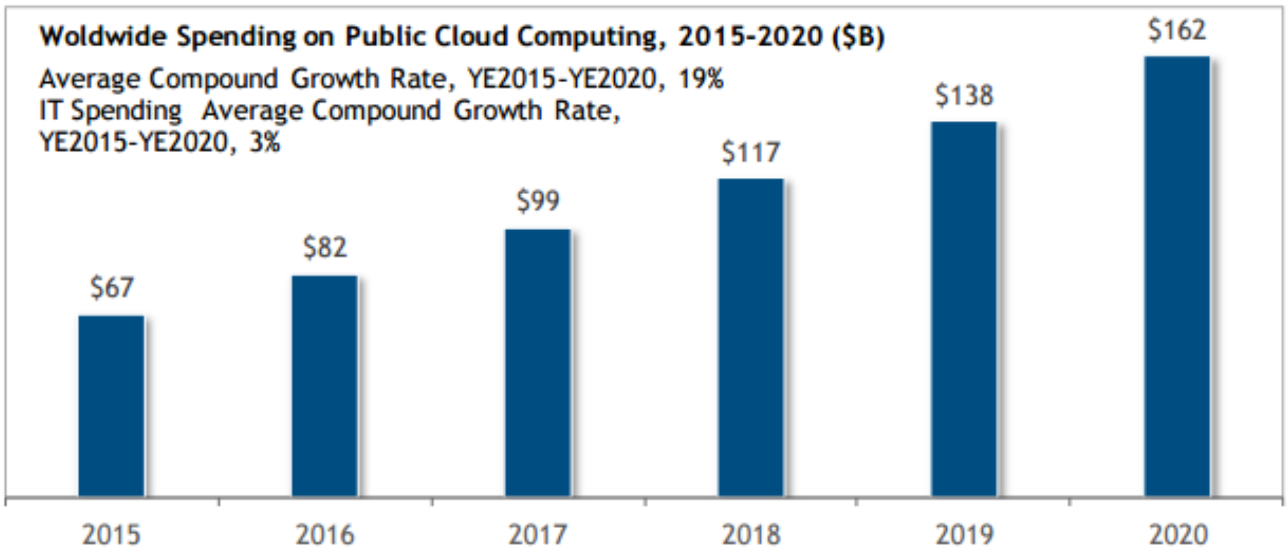
中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Outline

- Background and motivation
- Problem
- Framework
- Evaluation
- Discussion

The Rapid Development Of Cloud Computing

The Rapid Growth of Cloud Computing, 2015-2020



Source: IDC, 2016

中国科学院计算技术研究所

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Concerns with Public Cloud

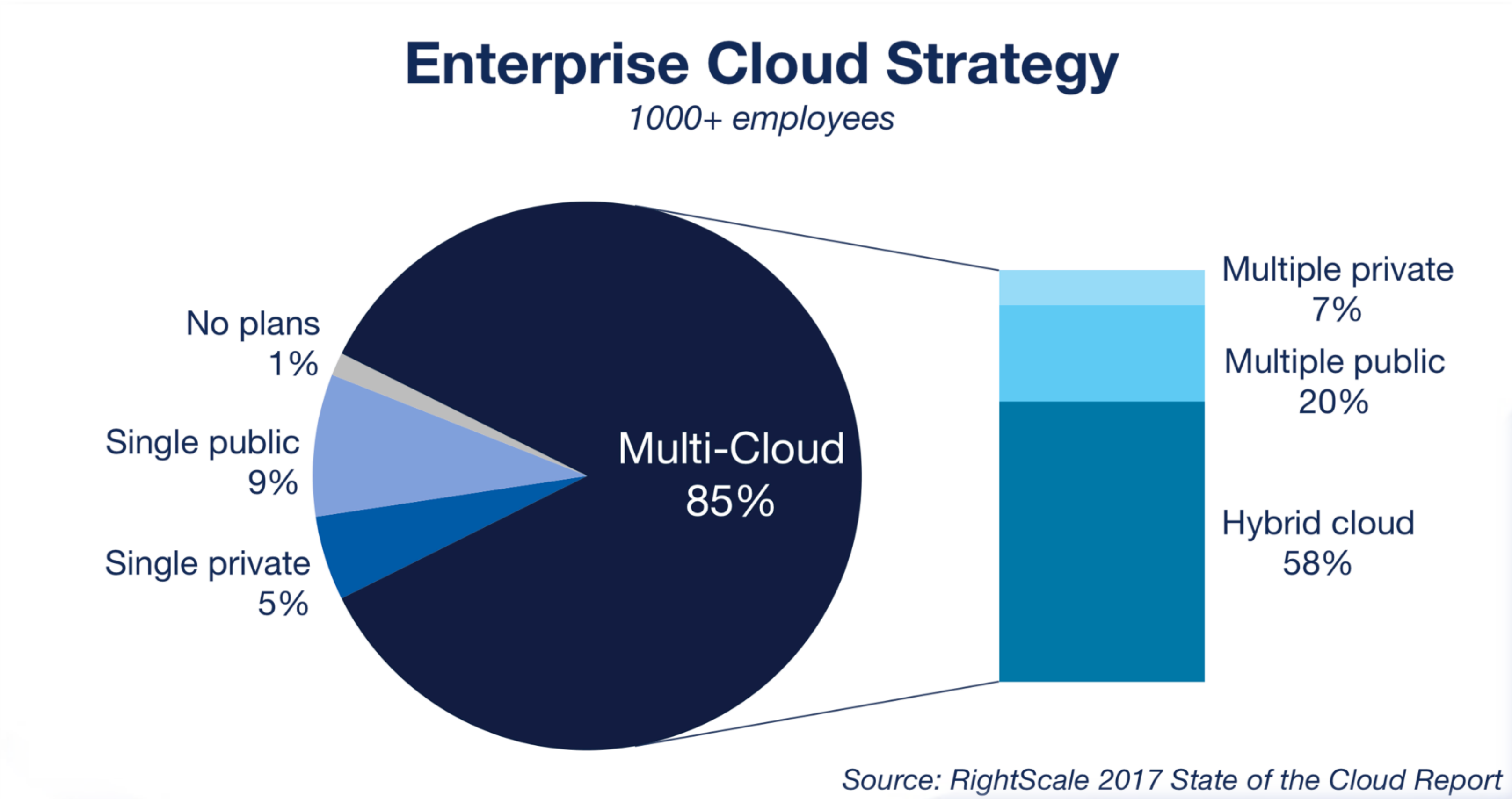
- Cheap cost

- Scalability

- SLA

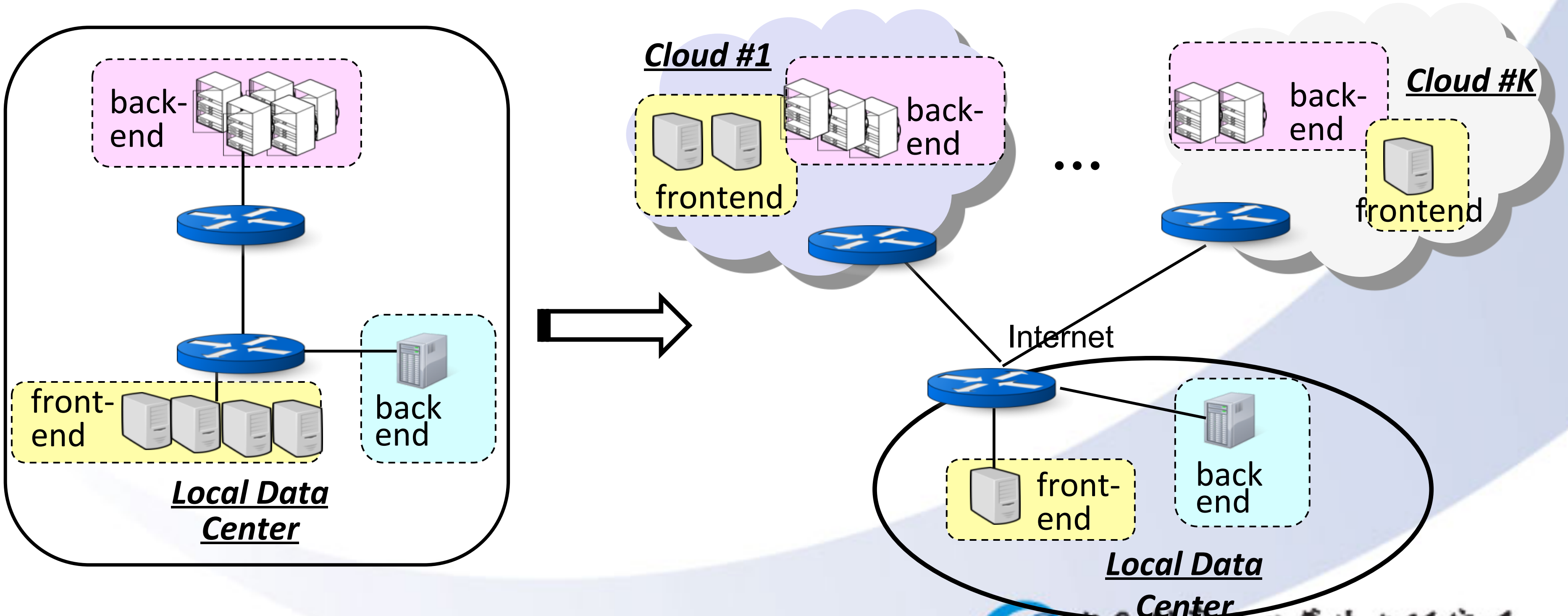
- Privacy

Call For Hybrid Cloud!

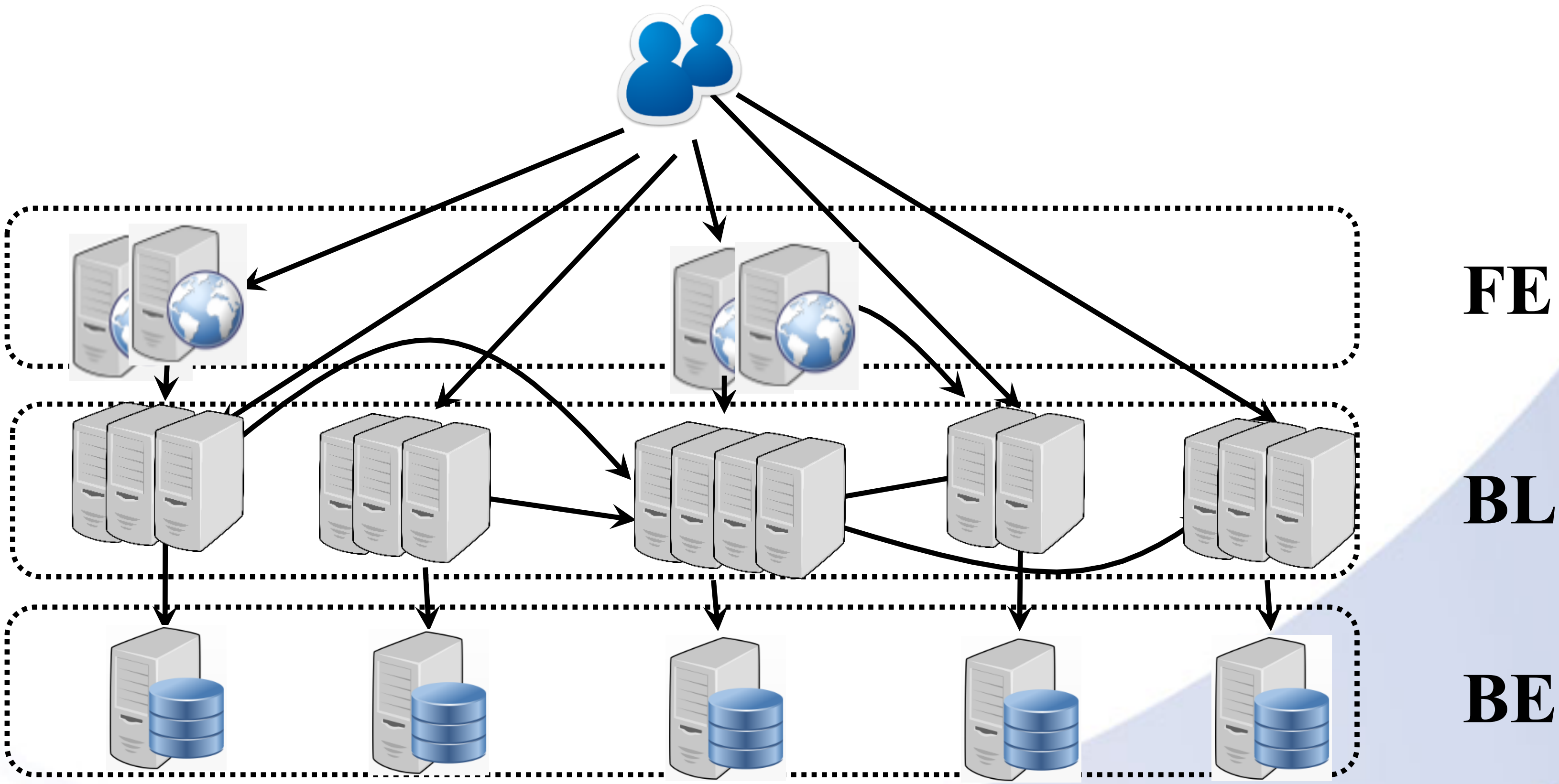


Our Focus: Planning Hybrid Cloud Layouts

- Cost savings, application response times, communication costs

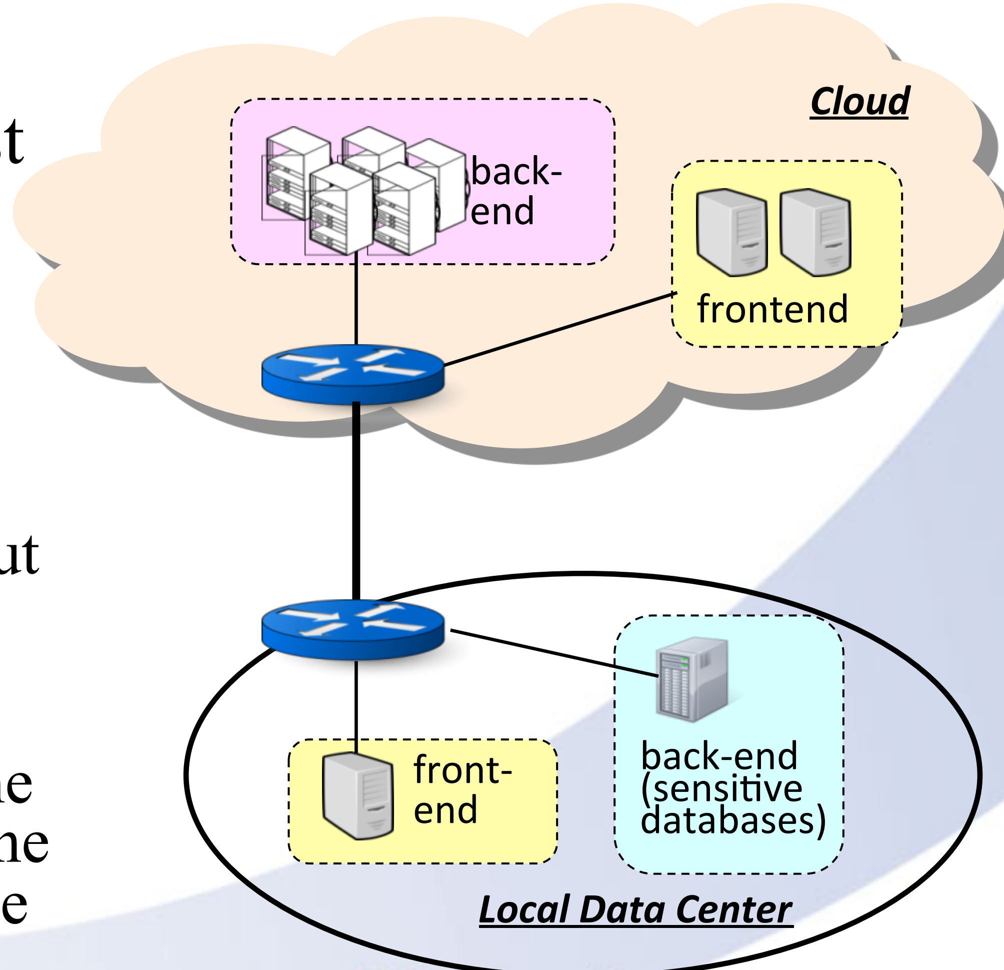


Model of Enterprise application



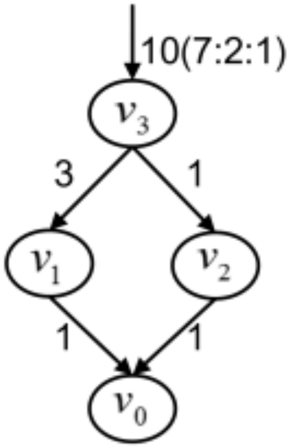
Abstracting the planning problem

- Objective: Maximize cost savings on migration
  - Benefits due to hosting servers in the cloud
  - Costs change related to wide area Internet communication (simple but practical linear model)
- Time Constraints:
  - The completion time of the application is defined as the maximum completion time of all workflows



中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Motivation Example



Plans	h_0	h_1	h_2	Time	Benefits	
1	All			25	0	A
2		All		30	40	NA
3			All	31	48	NA
4	v_1, v_3		v_0, v_2	28	20	A

Fig. 1. The application and four migration plans with the corresponding cost reduction and completion time.

Analysis

Complexity:

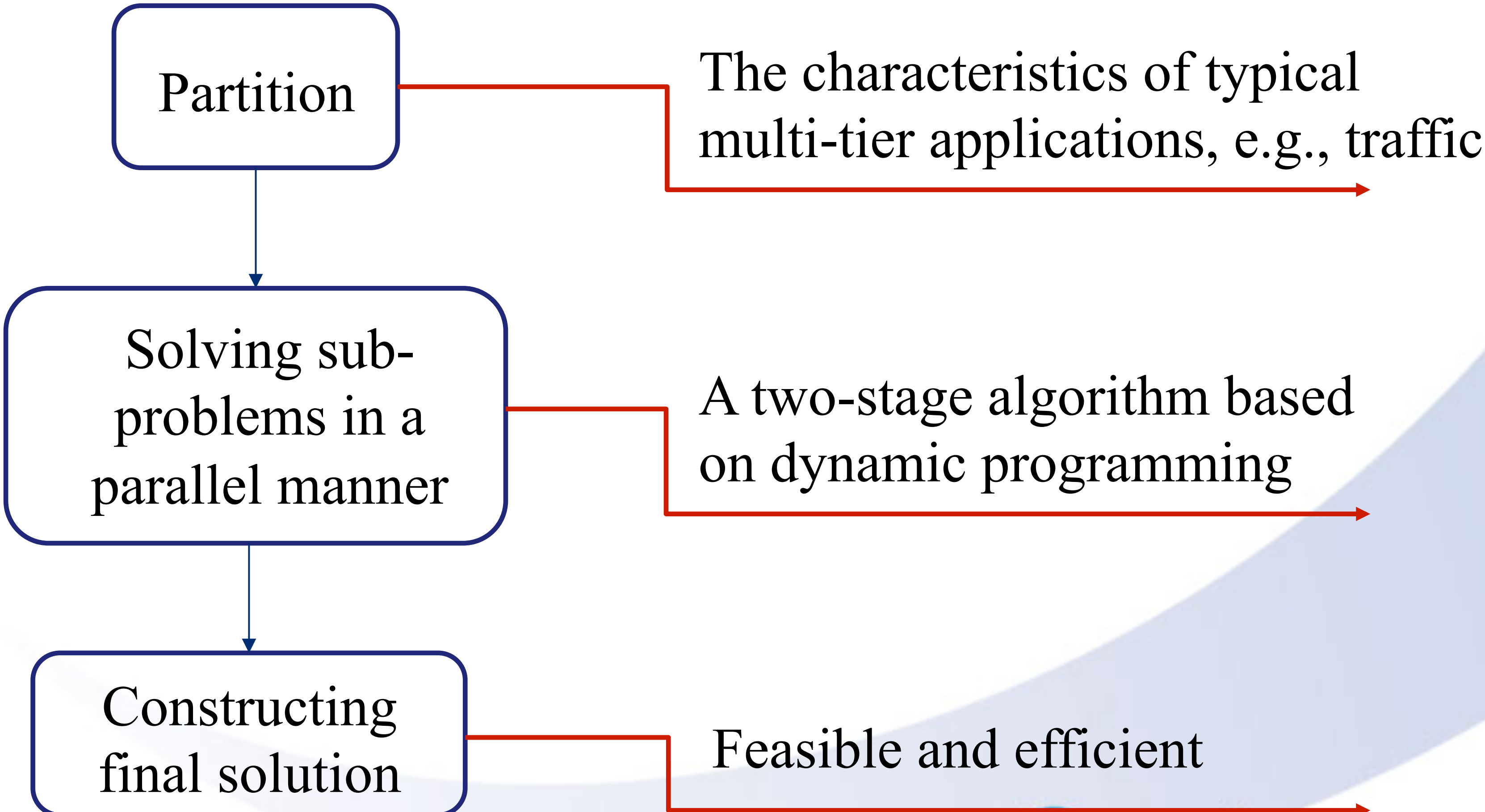
① Solving the general application deploy problem is NP-Complete.

Key observations:

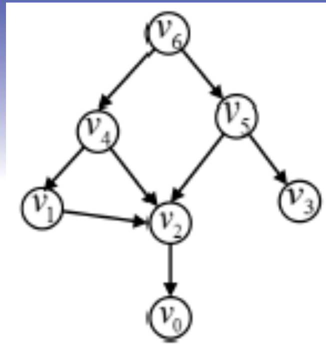
① Most of the multi-tier enterprise application can be easily divided into multiple DAGs.

② Solving the DAG deploy problem is much easier!

An overview of the Framework



A two-stage algorithm



- Step 1: transforming a DAG into a sequence

$$(v_0, v_2, v_1, v_3, v_4, v_5, v_6)$$

- Step 2: dynamic programming-based algorithm

c[i][j][k]: the max cost reduction of the subgraph rooted at node v_i when node v_i is assigned to cloud h_k and when the total delay is no larger than j.

$$c[i][j][k] = \begin{cases} \gamma_{i,k} + \sum_{1 \le h \le S_i} \{\max_{1 \le k_{ih} \le M} (c[i_{ih}] \\ [j_{ih}][k_{ih}] - \beta_{i,i_{ih},k,k_{ih}})\}, \text{if } \max_{1 \le h \le S_i} \\ (j_{ih} + d_i^e + d_{i,k}^a + d_{i,i_{ih},k,k_{ih}}) \le j; \\ \text{no feasible solution, otherwise.} \end{cases}$$

- Case 1: all the children of node v_i has only one parent node
- Case 2: at least one of the children of node v_i has multiple parent nodes

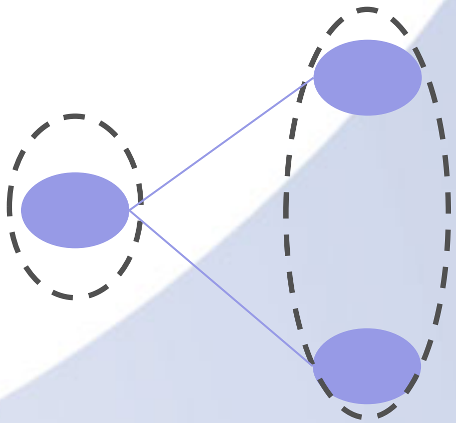
The evaluation setting

Applications:

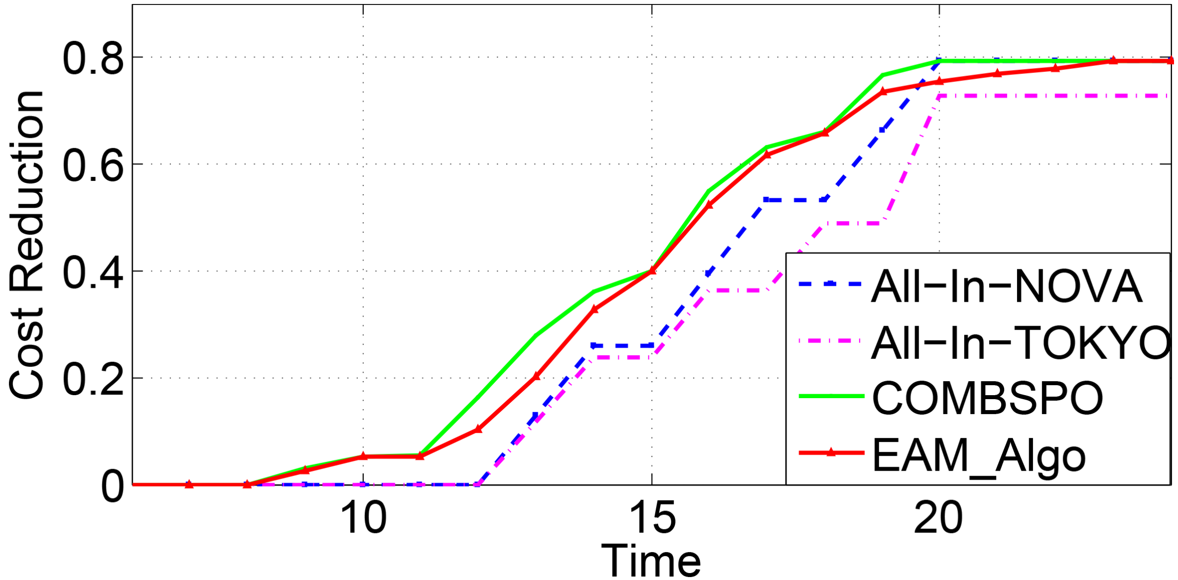
① Six randomly generated DAGs as application architectures.
  ① Each DAG involves a number of nodes between 500 to 2K.
  ② Each communication pair is associated with unit traffic.
  ③ Simulation results are

Hybrid cloud:

① A local cloud.
② Two public cloud from amazon:
  ① one in Northern Virginia (NOVA).
  ② One in Tokyo.

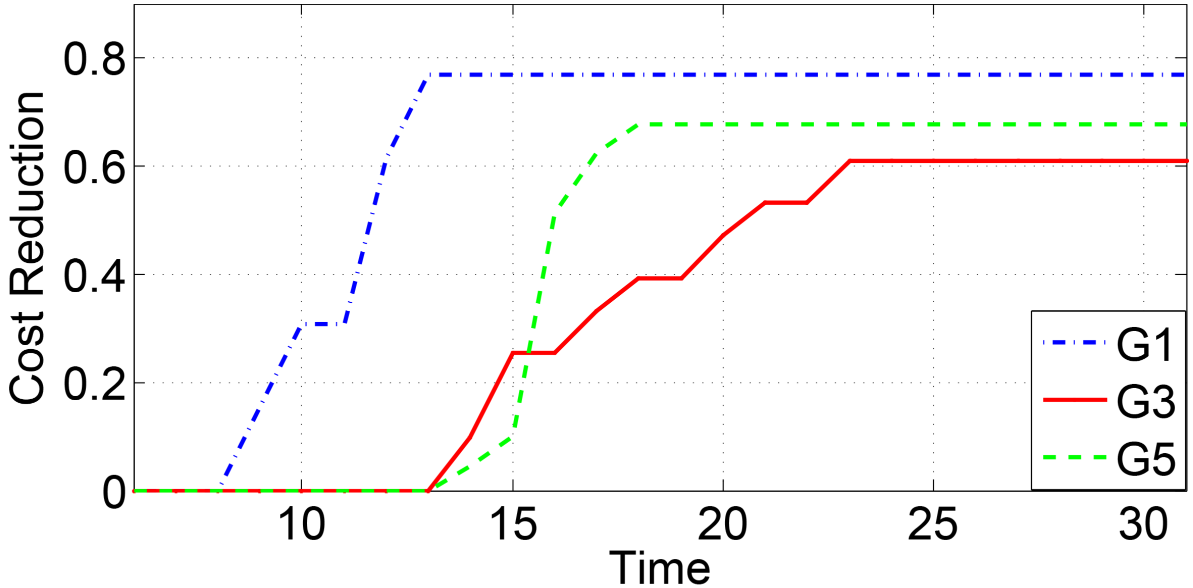


The performance-cost reduction



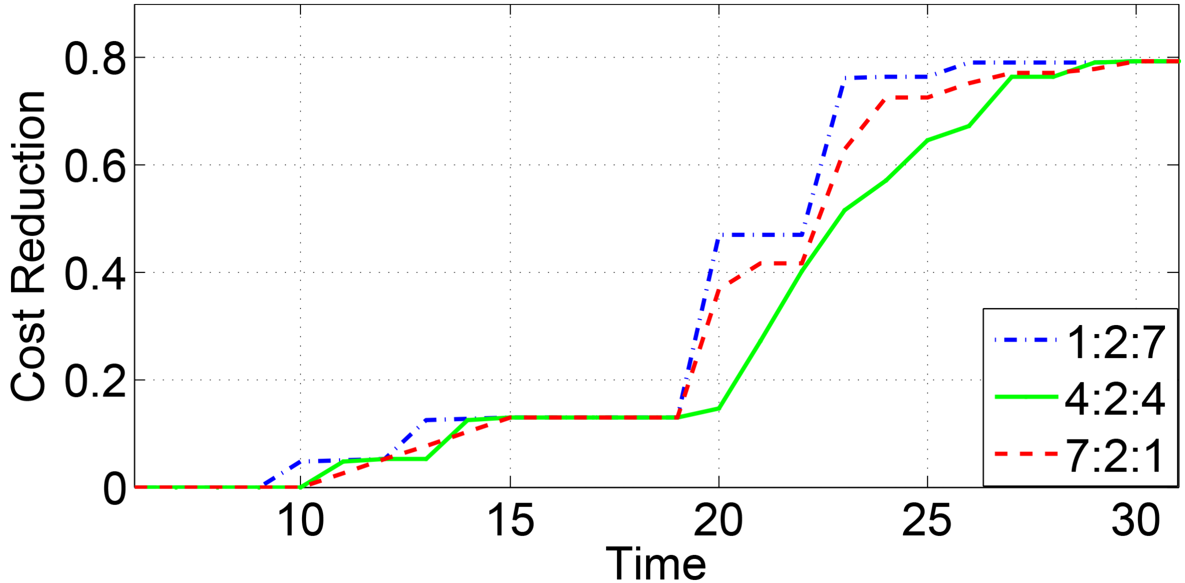
- (a) The average cost reduction of different strategies under different time constraints.
- (b) our framework can bring up to 79.15% cost reduction to enterprises. Besides, the cost reduction obtained by our algorithm is close to that of the optimal solution solved by COMBSPO.
- (c) our framework performs better in reducing enterprise costs leveraging the hybrid cloud architecture under controllable time overhead than the other two strategies.

The performance-the benefit-time tradeoff



- (a) by varying the value of time constraint, one can obtain a large cost reduction with large time overheads.
- (b) Choosing a proper value for an application depends on the performance requirement of the application manager.

The performance-the effect of user location.



- (a) migrating applications that have larger percentage of external user to cloud will bring more cost reduction than migrating the ones that have smaller percentage of external user.
- (b) The cost reduction of migrating the applications with users evenly distributed in three regions is the least.

Thanks for your attention!