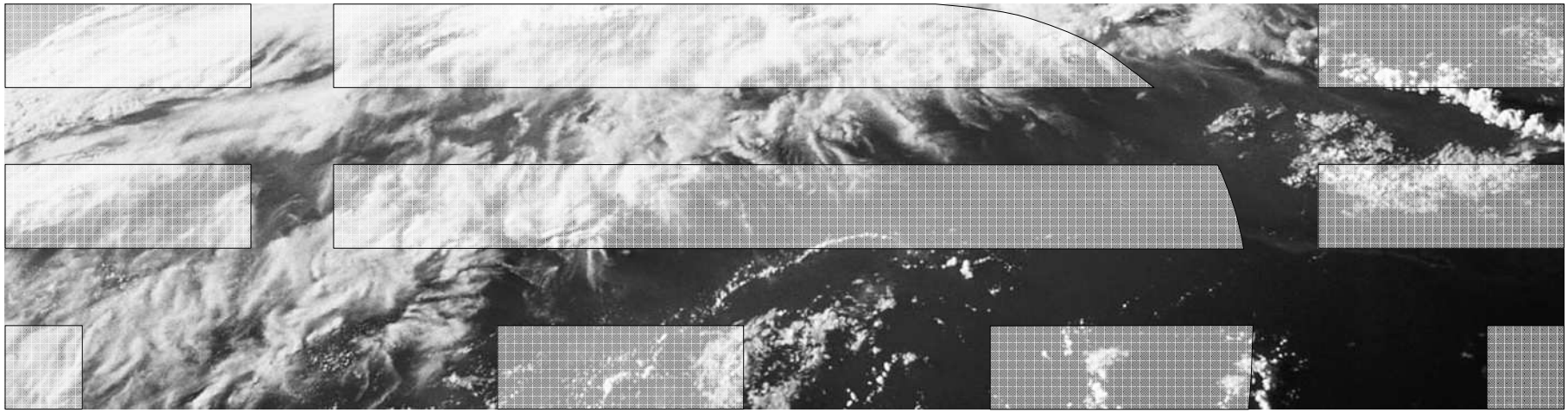
Kay Sripanidkulchai, Sambit Sahu, Yaoping Ruan, Anees Shaikh, and Chitra Dorai
IBM T.J. Watson Research Center

Are Clouds Ready for Large Distributed Applications?

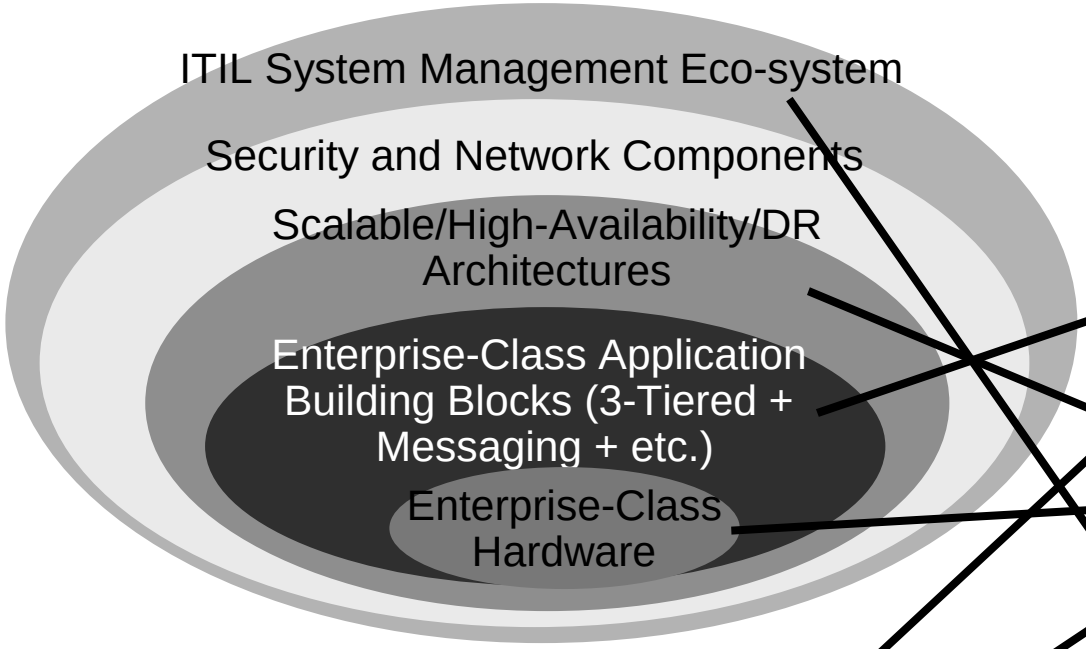


Outline

- What are users expecting from the cloud?
  - Establish a base-line for requirements
- Is the cloud meeting user requirements?
  - Service deployment
  - Service availability
  - Service problem resolution
- Where are opportunities?

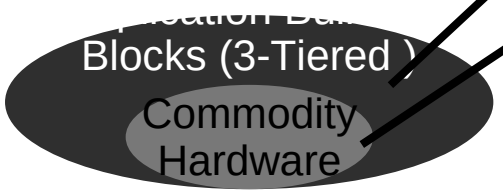
Enterprise vs. individual customers have different requirements

Typical Enterprise Application Architecture



Typical Small/Individual Application Architecture

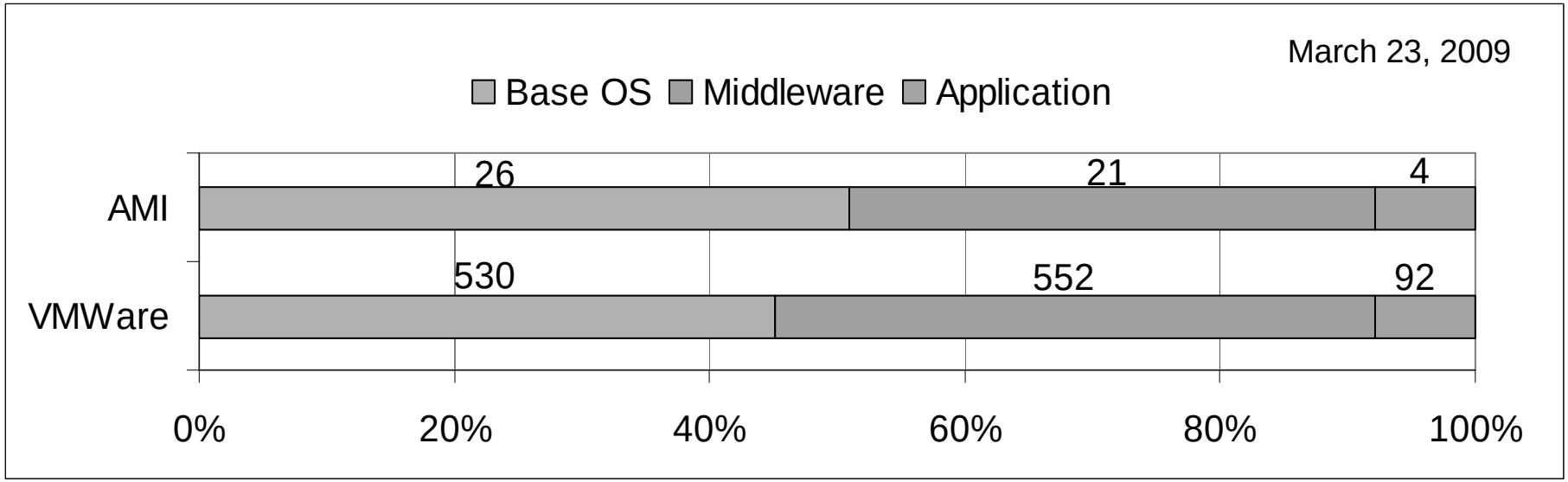
? ? ?



We study three primary requirements

- How to deploy large-scale distributed services on the cloud,
- How to deliver high availability services using clouds, and
- What to do when there are problems with services running on the cloud.
- For others, see [AFG et. al 08], [WSRV09]

Are there sufficient building blocks available to enterprise users to quickly deploy their services on the cloud?



Base OS and middle-ware images dominate the landscape. Where are the complex applications? Where are the multi-tier distributed applications with multiple images?

Towards supporting deployment of large-scale distributed applications….

- Service composition to support complex applications beyond single VMs.
  - Express relationships among these VMs denoting the dependencies at configuration time and at running time
  - Compose complex deployment from single and already built set of VMs, and
  - Instantiate the deployment based on the above stated dependencies.

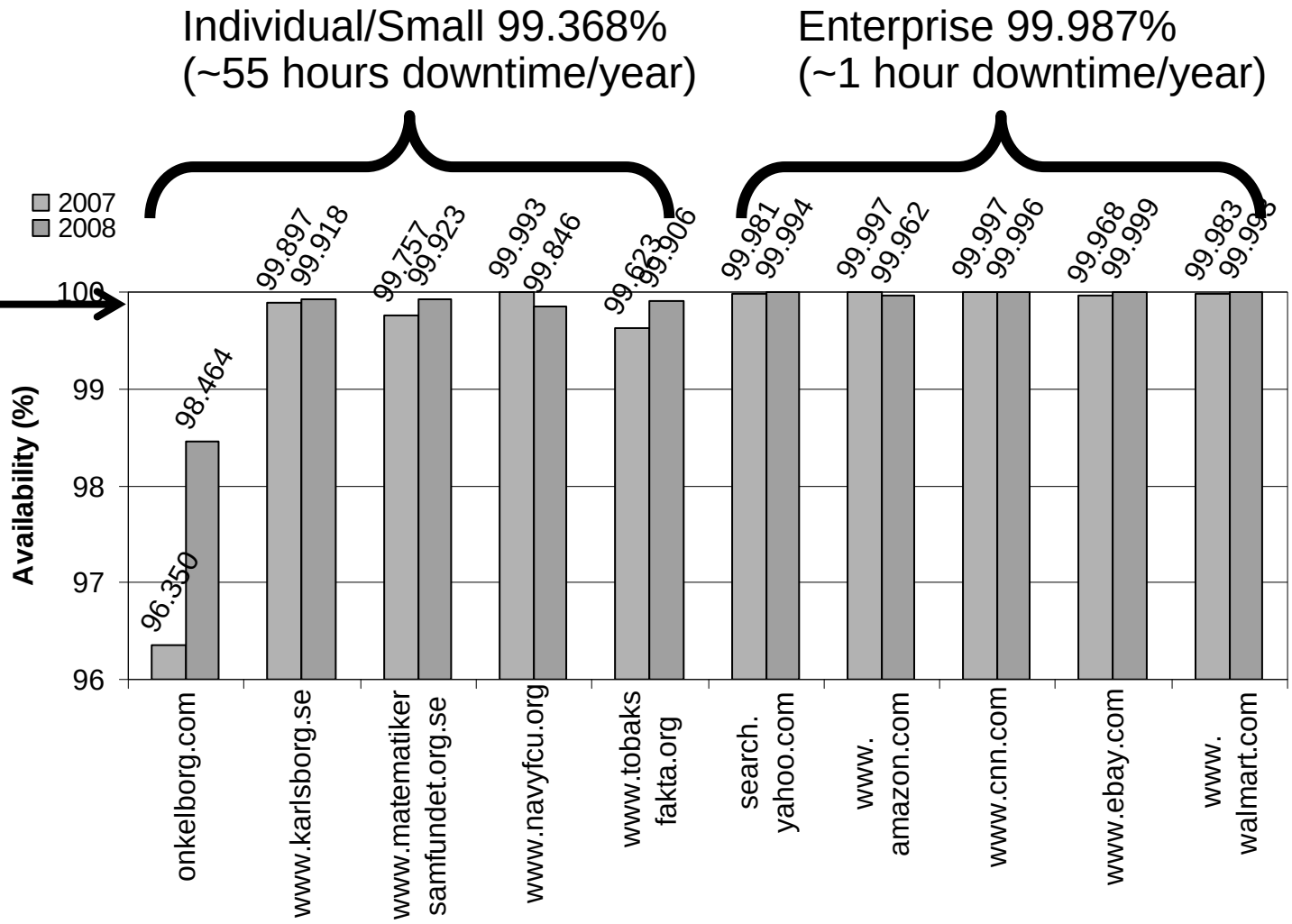
  Current status: Already headed this way with third-party services such as 3Tera and RightScale, but will eventually need a common standard.

- Transformation of existing enterprise service deployment into a cloud-based deployment
  - Discovery of application configuration and dependency of the enterprise services to be migrated to the cloud
  - Determine the amount of infrastructure resources needed on the cloud and map application components to the resources
  - Support for provisioning the service and migrating to the cloud in an easy and quick manner, without incurring service down time. Can we do this live?

  Current status: Discovery techniques and dependency graphs have been explored in other contexts such as problem determination. The rest is open.

There are gaps in service availability requirements for enterprise users

State-of-the-art cloud SLA at 99.95% or ~4 hours downtime/ year.



Bridging the gap in service availability requirements

- Implementing scaling architectures in the cloud
  - Templates and rules to determine based on system conditions to automatically leverage the appropriate architectural solution
  - Commoditize the expertise so that it can be reused by different cloud users

  Current status: components such as content delivery networks, load-balancing and automatic scaling (elasticity) are available, but best practices for how to use these components have not been established. Can the cloud just automatically do this for me?

- Extending availability beyond one cloud
  - API or framework to commoditize the construction of high availability services delivered across multiple clouds

  Current status: few service providers -- too early but already concerned about lock-in

- Using the latest and greatest virtualization capabilities
  - Live migration to avoid down time

  Current status: non-existent inside one cloud and across clouds. Who gets to decide when/why to migrate? The user or the cloud provider?

Best practice in service problem resolution faces scaling challenges

Feature Request	HowTo/ Info	Problem		
		Cloud Error	User Error	Unknown
10%	56%	25%	64%	11%

Amazon EC2 Forum: April 1-7, 2009

Observations

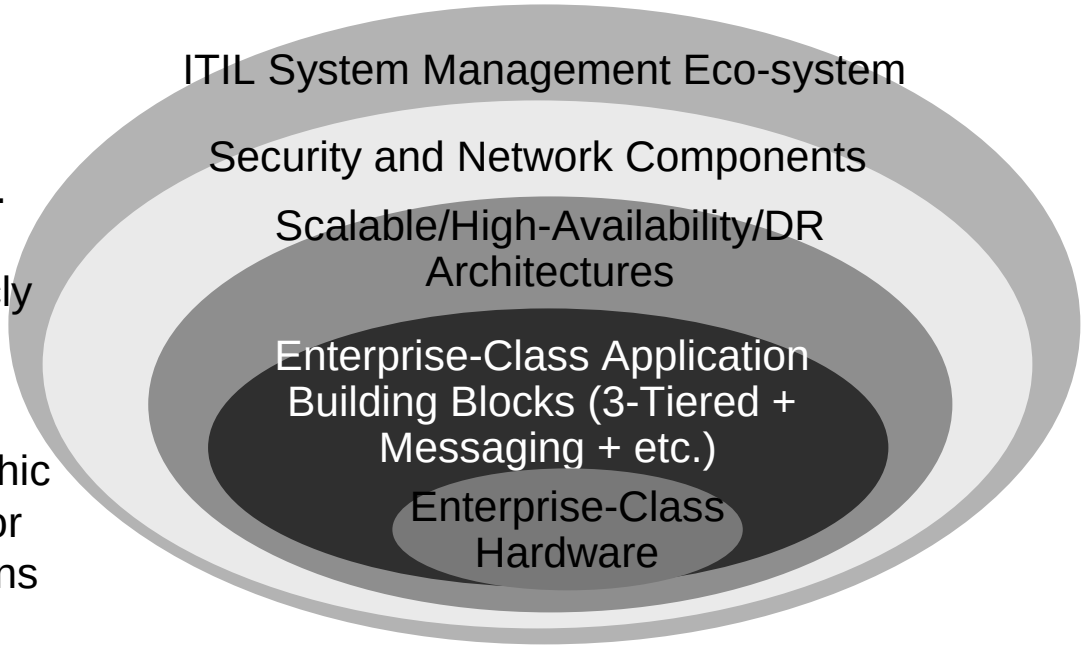
- Top problems: Instance, EBS, Security
- The same symptom presented to the user has many underlying root causes
- Resolution process is highly manual and ad-hoc; manual information sharing is error-prone and not scalable
- Users do not know what is happening in the underlying infrastructure and cloud provider does not know what happening in the users applications

Where to go next

- Define an API for information sharing between users and providers that addresses privacy concerns
  - Is a minimum of a binary “your problem” vs. “my problem” query sufficient?
  - Can all of a user’s instances be managed together?

Summary

- Explored three requirements from the perspective of cloud users
  - Compared individual/small users vs. enterprise users
  - Established a base-line using publicly available data
- Service deployment
  - Current practice focuses on monolithic systems, with some initial support for more complex distributed applications underway.
  - Future work to support large-scale distributed architectures is needed.



- Service availability
  - SLA's are in place and high enough to meet individuals' needs.
  - Future work to increase availability is crucial to attract enterprise users and would also benefit individual users.
- Problem resolution
  - Current manual process faces scaling challenges
  - Future work to reduce the load on the cloud support staff such as providing cloud users with enough *visibility* into the cloud infrastructure to independently identify the root cause of problems is needed to scale up.