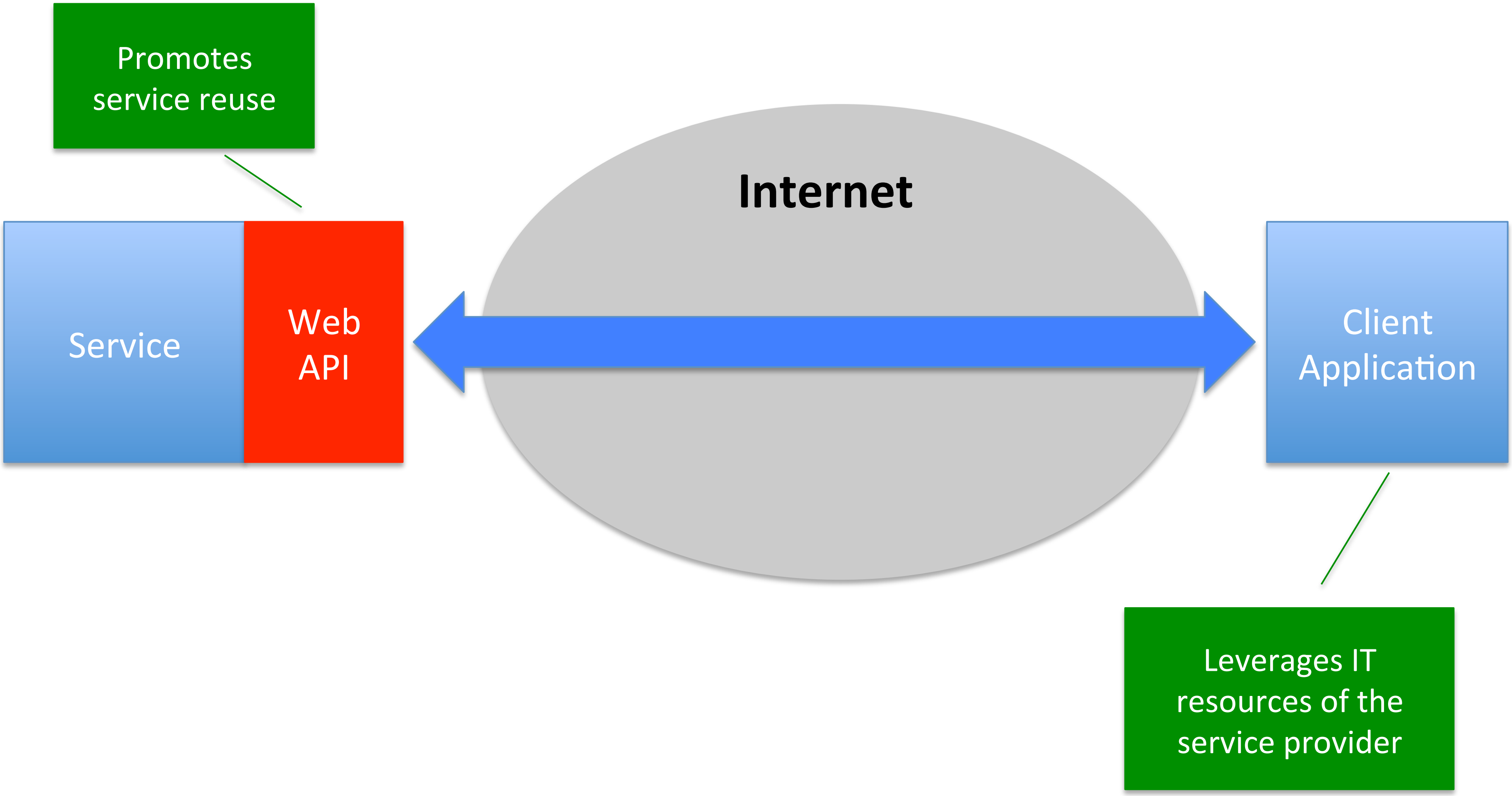
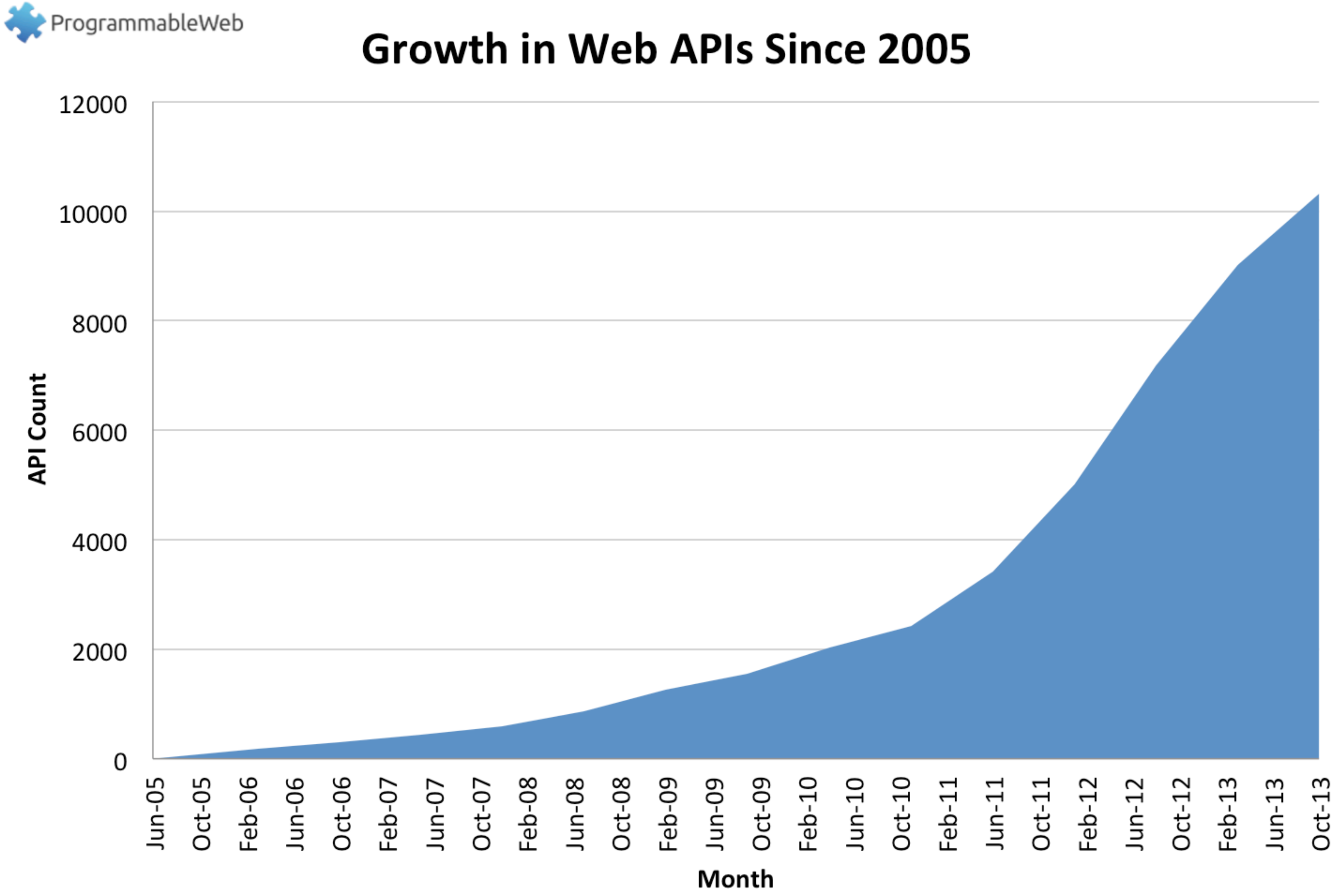
Response Time SLAs for Cloud-hosted Web Applications

Hiranya Jayathilaka
Prof. Chandra Krintz
Prof. Rich Wolski
Computer Science Dept., UC Santa Barbara

SOCC 2015

Web APIs


Promotes service reuse
Internet
Service
Web API
Client Application
Leverages IT resources of the service provider

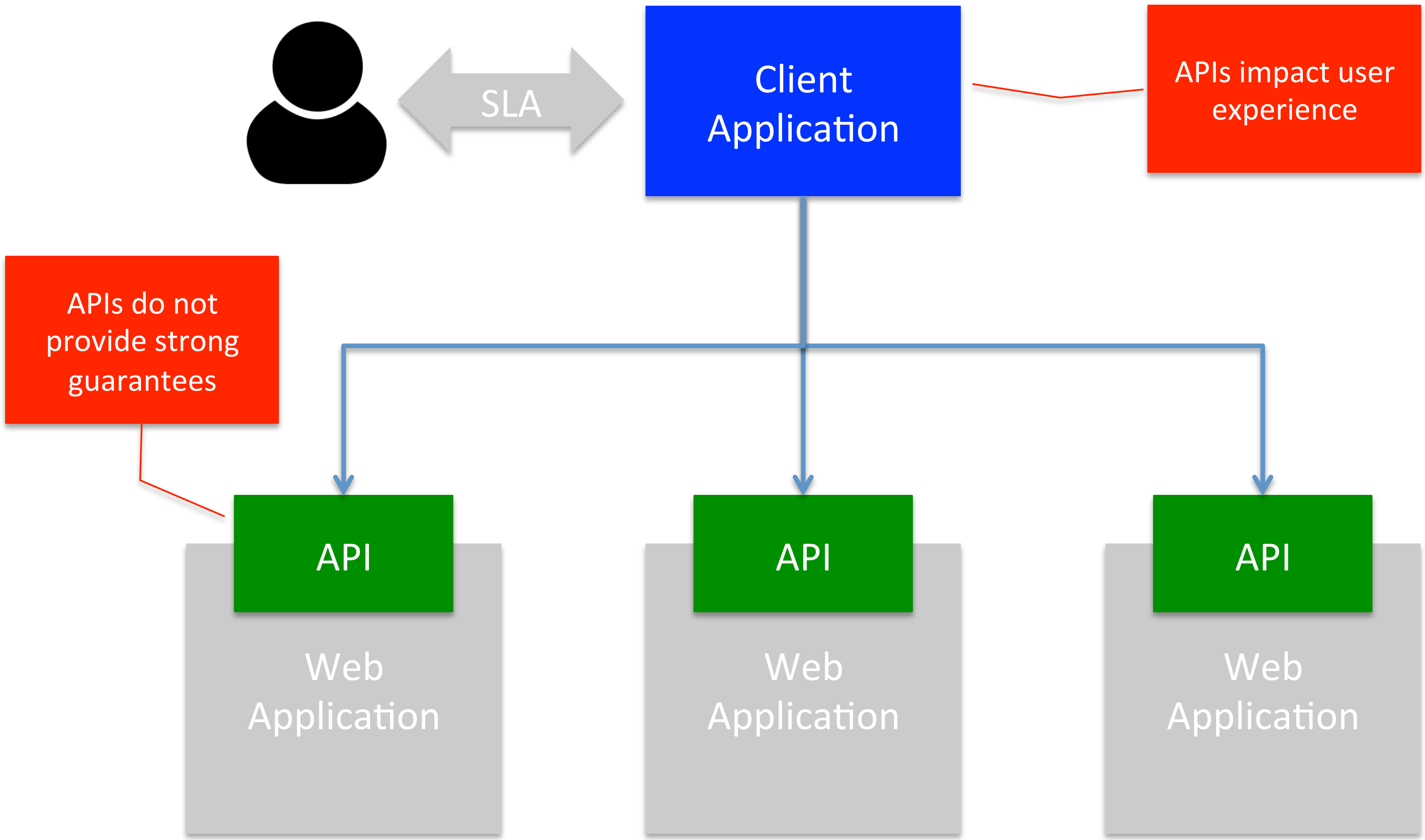


Growth in Web APIs Since 2005

Number of API Today: ~14,000

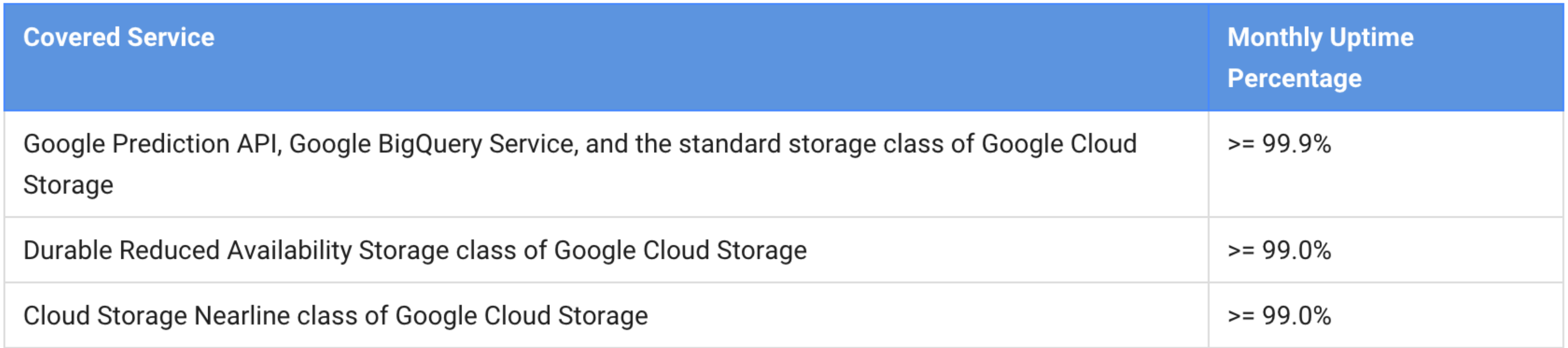
Source: http://www.programmableweb.com/api-research

Web APIs are Now IT Resources



Application SLAs and "The Cloud"

- Modern cloud platforms only provide *some* uptime SLAs for individual APIs



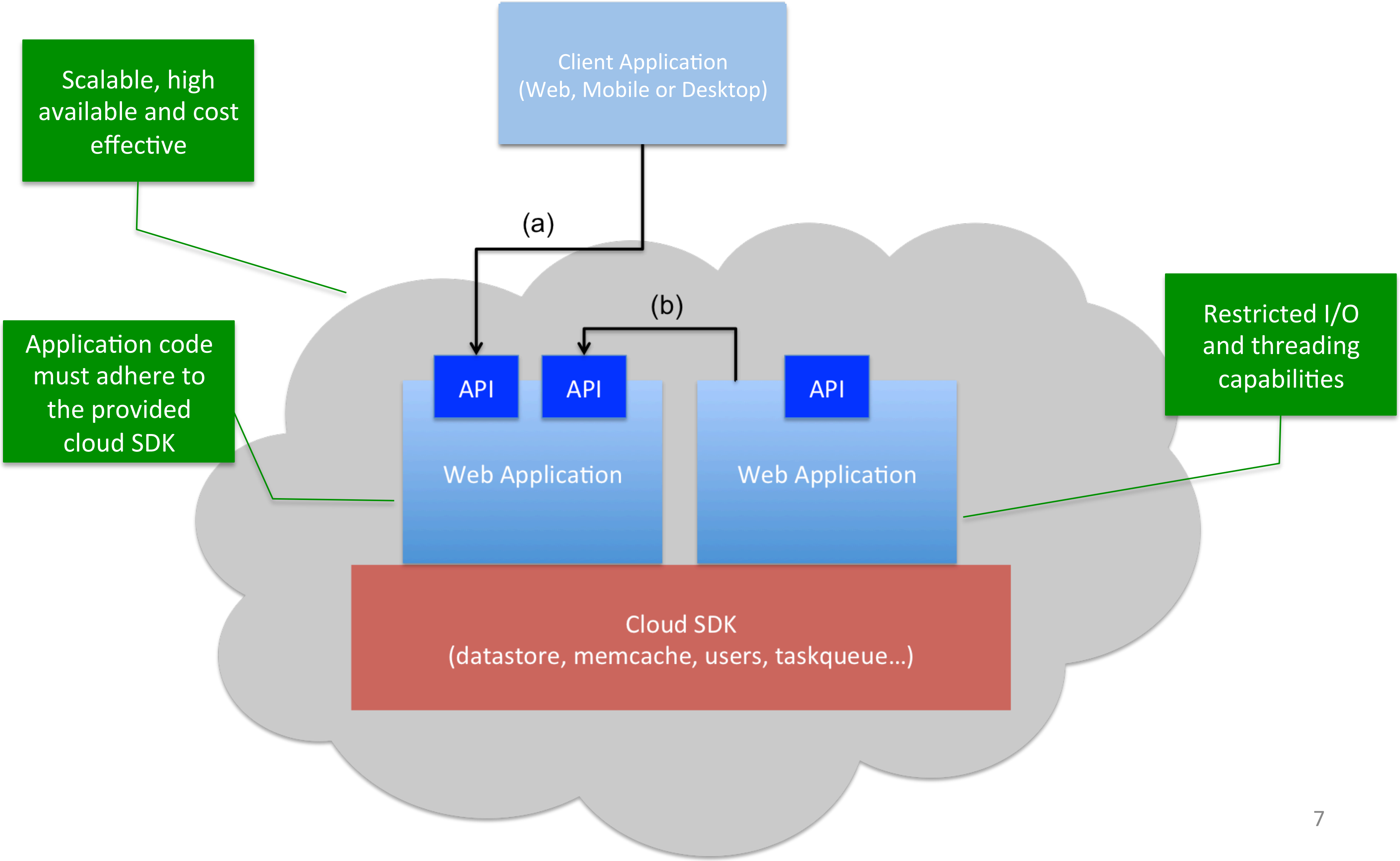
Covered Service	Monthly Uptime Percentage
Google Prediction API, Google BigQuery Service, and the standard storage class of Google Cloud Storage	>= 99.9%
Durable Reduced Availability Storage class of Google Cloud Storage	>= 99.0%
Cloud Storage Nearline class of Google Cloud Storage	>= 99.0%

- Cloud platforms do not provide SLAs on deployed user applications and APIs.

Performance SLAs in the Cloud

- ***Question:*** Is it possible to determine, automatically, performance SLAs for cloud-hosted applications and APIs?
- ***Our solution:*** Cerebro
  - Predicts the response time of future web-API invocations from historical measurements
  - Fully automatic
  - For PaaS clouds

PaaS Clouds for Web Services


Scalable, high available and cost effective
Application code must adhere to the provided cloud SDK
Client Application (Web, Mobile or Desktop)
(a)
(b)
API
API
API
Web Application
Web Application
Restricted I/O and threading capabilities
Cloud SDK (datastore, memcache, users, taskqueue…)

PaaS Client Application Survey

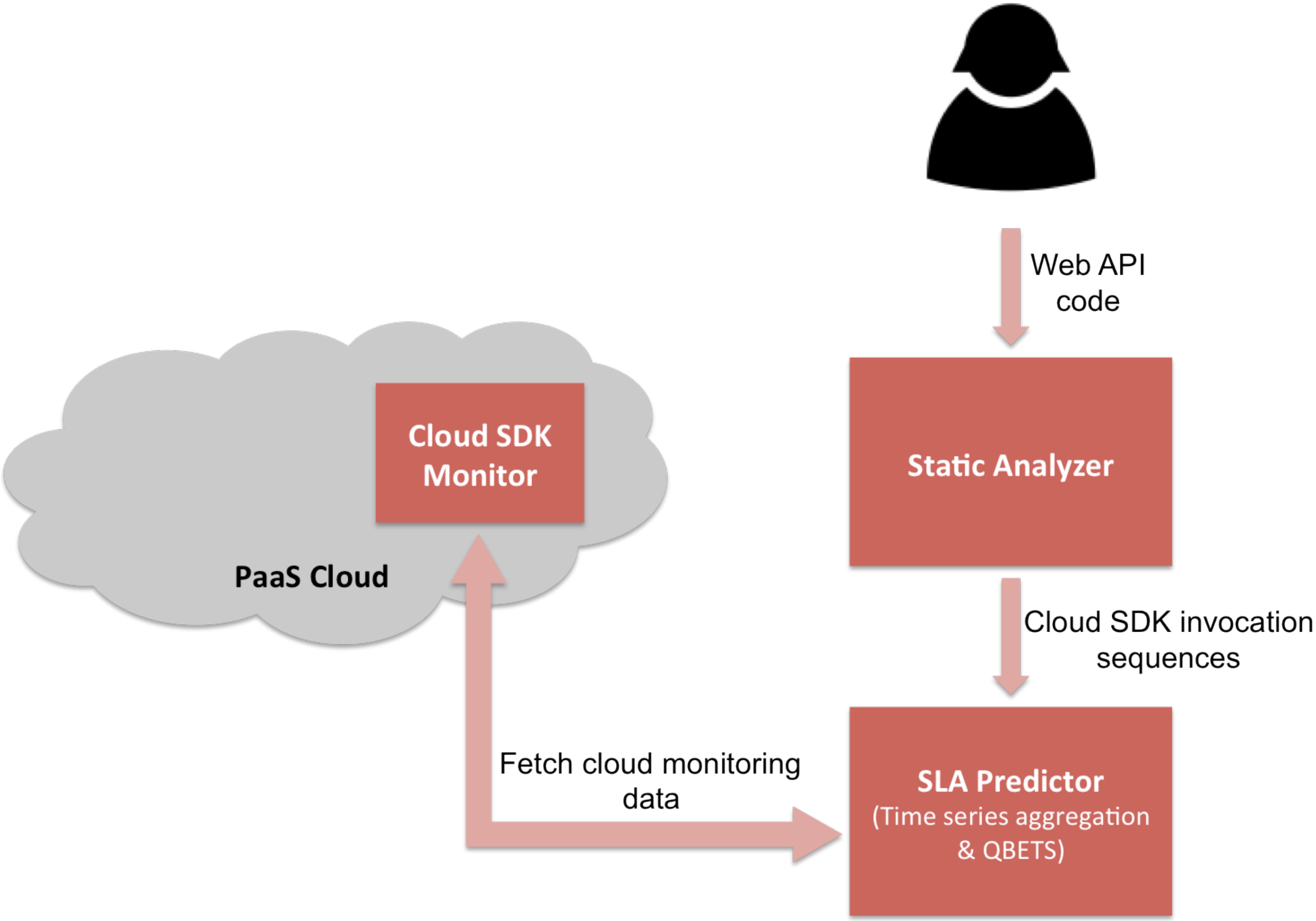
PaaS Client Applications…

- Don’t have many branches
  - 99% of the methods have < 36 paths
- Don’t have many loops
  - 88% of the methods have no loops
- Spend most of their time executing cloud SDK calls (> 94%)

So…

- PaaS applications are highly amenable to static analysis
- Cloud SDK calls essentially define client-perceived application performance

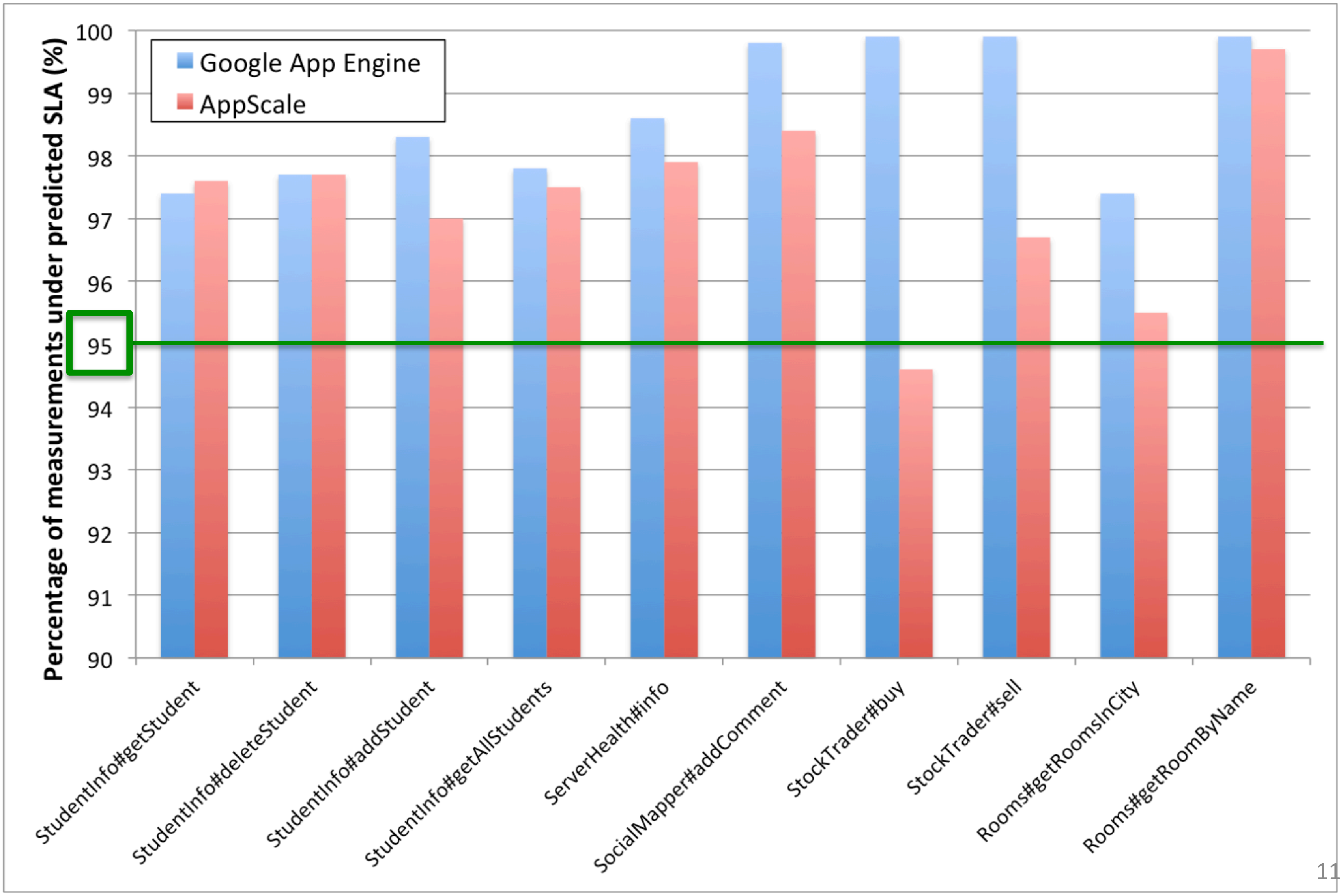
Cerebro Architecture


Web API code
Cloud SDK Monitor
PaaS Cloud
Static Analyzer
Cloud SDK invocation sequences
Fetch cloud monitoring data
SLA Predictor
(Time series aggregation & QBETS)

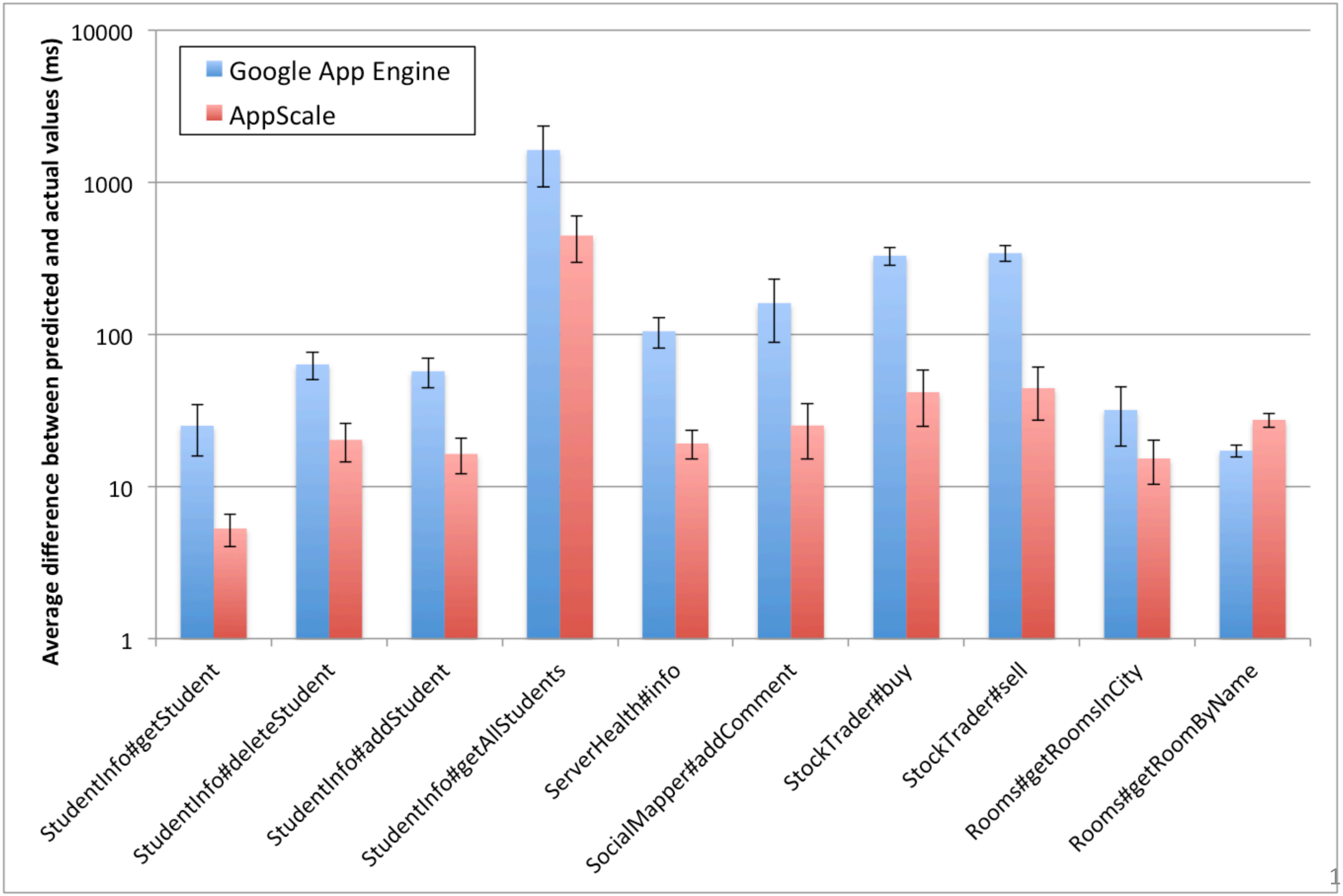
QBETS: Queue Bounds Estimation from Time Series

- Analyzes the first *n* entries in a time series
- Predicts an upper bound for the *(n+1)*th entry
  - $QBETS([x_1,x_2,\ldots x_n], p) = Q$ where $p \in (0,1)$
  - $P(x_{n+1} \leq Q) \geq p$
- Cerebro uses QBETS to predict response time SLAs of the form:
  - Operation *O* responds *under T* milliseconds (100*p)* *%* of the time

Evaluation: Prediction Correctness



Evaluation: Prediction Tightness



Conclusions and Future Work

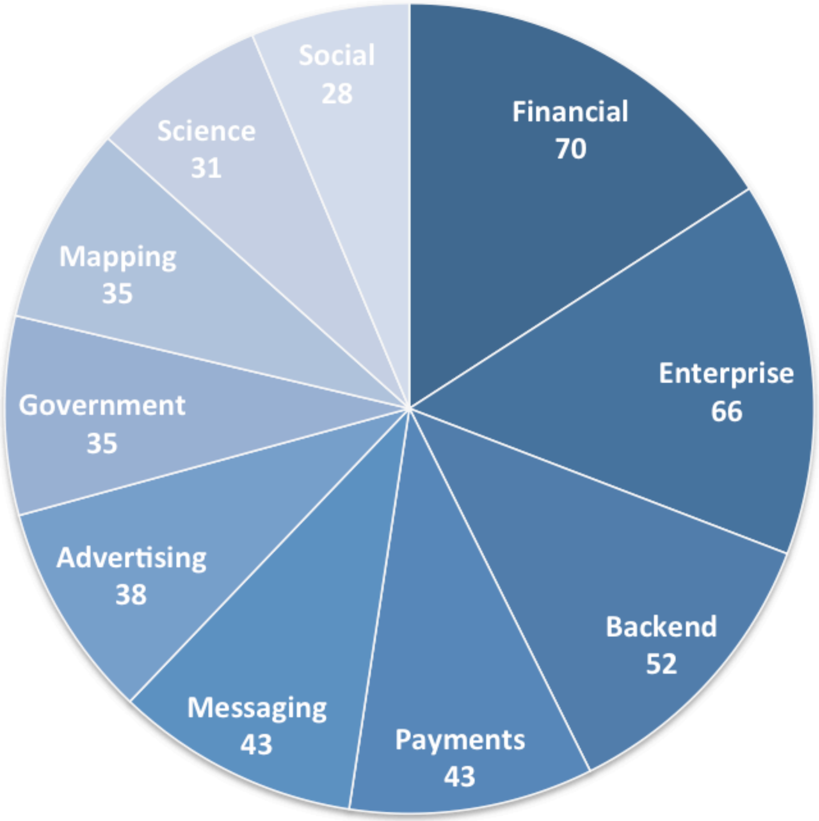
- Cerebro predictions are correct and moderately tight
  - Necessary conditions for use in an SLA
- SLA durability period analysis
  - GAE: 26.8 hours
  - AppScale: 33.7 hours
- SLA-related policy enforcement at deployment time with EAGER

Thank You! Questions?

- Hiranya Jayathilaka (hiranya@cs.ucsb.edu)
- The UCSB Lab for Research on Adaptive Computing Environments (RACELab)
  - http://www.cs.ucsb.edu/~ckrintz/racelab.html

ProgrammableWeb

Fastest Growing Web API Categories
(6 Months)



Non-commercial entities are joining the API party too…

- White House API Program: https://www.whitehouse.gov/digitalgov/apis
- IEEE APIs: http://ieeexplore.ieee.org/gateway/
- UC Berkeley APIs: https://api-central.berkeley.edu/

Prototype and Experiments

- SDK monitor: App Engine Java app
- Tests conducted on Google App Engine public cloud, and AppScale private cloud (running on a 4-node Eucalyptus cluster)
- Network delay between client and Google is included but not modeled or predicted explicitly