CSC2231: Making clusters fault-tolerant

http://www.cs.toronto.edu/~stefan/courses/csc2231/05au

Stefan Saroiu

Department of Computer Science

University of Toronto

Administrivia

- **Project proposals due in 1 week (noon Thursday)**
  - Create Web page with brief project proposal (HTML,TXT)
    - What is the problem you are solving?
    - Why is the problem interesting?
    - Why is the problem hard?
    - How are you planning to solve the problem?
    - What is the related work?

What's all about these 9's?

- **two 9's ~ 3.5 days per year**
- **three 9's ~ 10 hours per year**
- **four 9's ~ 1 hour per year**
- **five 9's ~ 5 mins per year**
- **six 9's ~ 30 secs per year**
- **seven 9's ~ 3 secs per year**

What's all about these 9's?

- **two 9's ~ 3.5 days per year**
- **three 9's ~ 10 hours per year**
- **four 9's ~ 1 hour per year**
- **five 9's ~ 5 mins per year** **nuclear reactor monitoring**
- **six 9's ~ 30 secs per year**
- **seven 9's ~ 3 secs per year**

What's all about these 9's?

- **two 9's ~ 3.5 days per year**
- **three 9's ~ 10 hours per year**
- **four 9's ~ 1 hour per year**
- **five 9's ~ 5 mins per year** **nuclear reactor monitoring**
- **six 9's ~ 30 secs per year** **telephone switches**
- **seven 9's ~ 3 secs per year**

What's all about these 9's?

- **two 9's ~ 3.5 days per year**
- **three 9's ~ 10 hours per year**
- **four 9's ~ 1 hour per year**
- **five 9's ~ 5 mins per year nuclear reactor monitoring**
- **six 9's ~ 30 secs per year telephone switches**
- **seven 9's ~ 3 secs per year**
- **..**
- **nine 9's ~ 30 ms per year in-flight computers**

What's our target?

- **How available should (Amazon|Google|eBay)'s clusters be?**

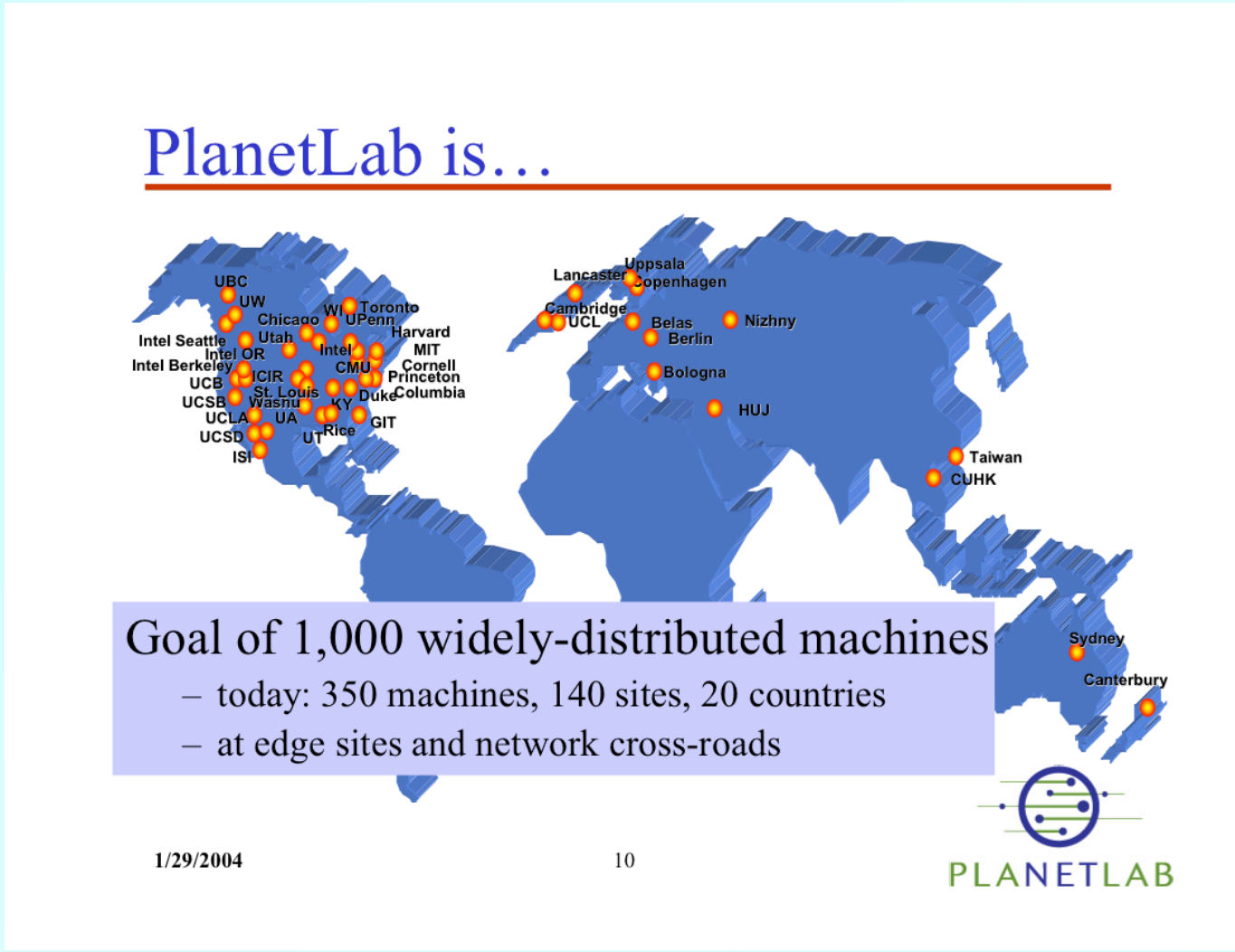
What's our target?

- **How available should (Amazon|Google|eBay)'s clusters be?**
  - Not more available than the availability of Internet paths
  - Not less available than Internet users' timeout
    - "reload consistency" ==> O(2-3s)
  - Not less available than the competitors' availability

Cost of Downtime

- **It is easy to translate availability to lost $$$**
  - Cost of 1 hour downtime = average revenue per hour + employee costs per hour
  - Hidden costs:
    - Customers' retention rates
    - Comparative costs relative to the rest of the industry
- **Internet service availability is a function of:**
  - Internet routing availability
    - BGP routing layer is known for slow fail-over
    - Little is known about ISPs failures
  - Cluster's availability

PlanetLab


PlanetLab is…
UBC
UW
Chicago
Utah
Intel Seattle
Intel OR
Intel Berkeley
UCB
ICIR
UCSB
Wasnu
UCLA
UA
UCSD
ISI
St. Louis
UT
Rice
KY
GIT
WI
Toronto
UPenn
Harvard
Intel
MIT
CMU
Cornell
Princeton
Duke
Columbia
Lancaster
Uppsala
Copenhagen
Cambridge
UCL
Belas
Berlin
Nizhny
Bologna
HUJ
Taiwan
CUHK
Sydney
Canterbury
Goal of 1,000 widely-distributed machines
– today: 350 machines, 140 sites, 20 countries
– at edge sites and network cross-roads
1/29/2004
10
PLANETLAB

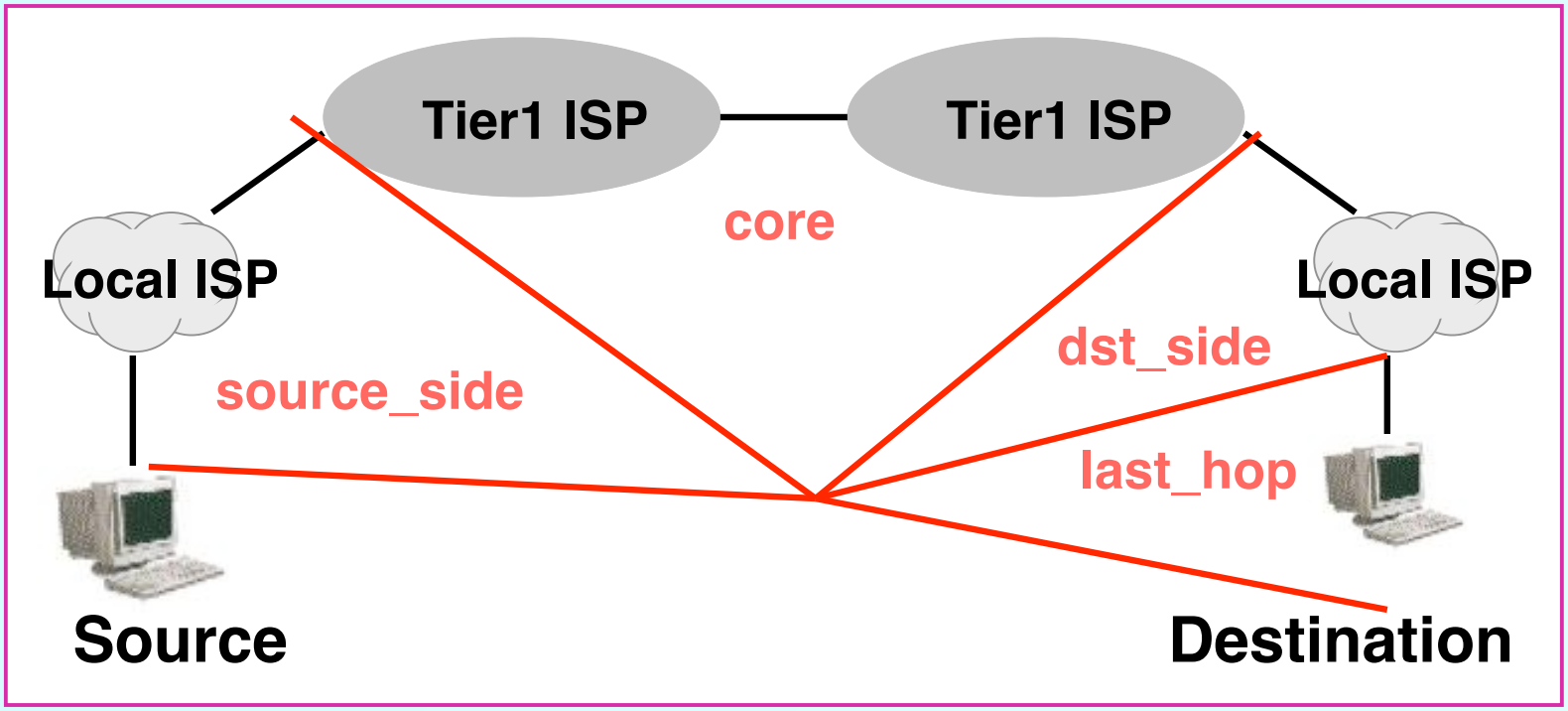
Internet Path Availability

The Availability of the Internet

- **Hard to quantify:**
  - What is a “representative” measurement?
  - How should degraded service be treated?

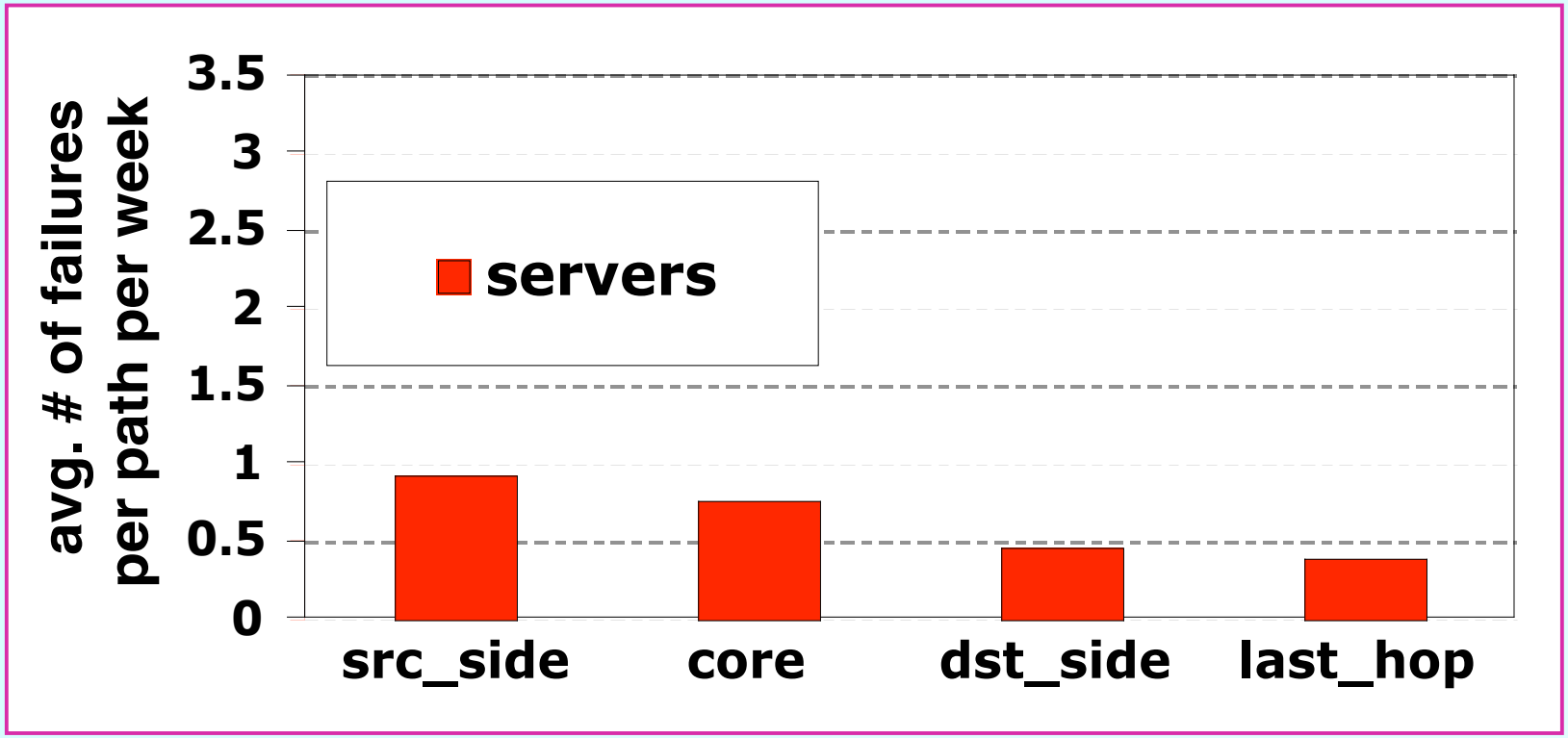
- **Wisdom says:**
  - Two-three 9’s
  - Hasn’t changed for almost a decade
  - Intra-domain routing is more reliable than inter-domain

Categories of Internet failure locations



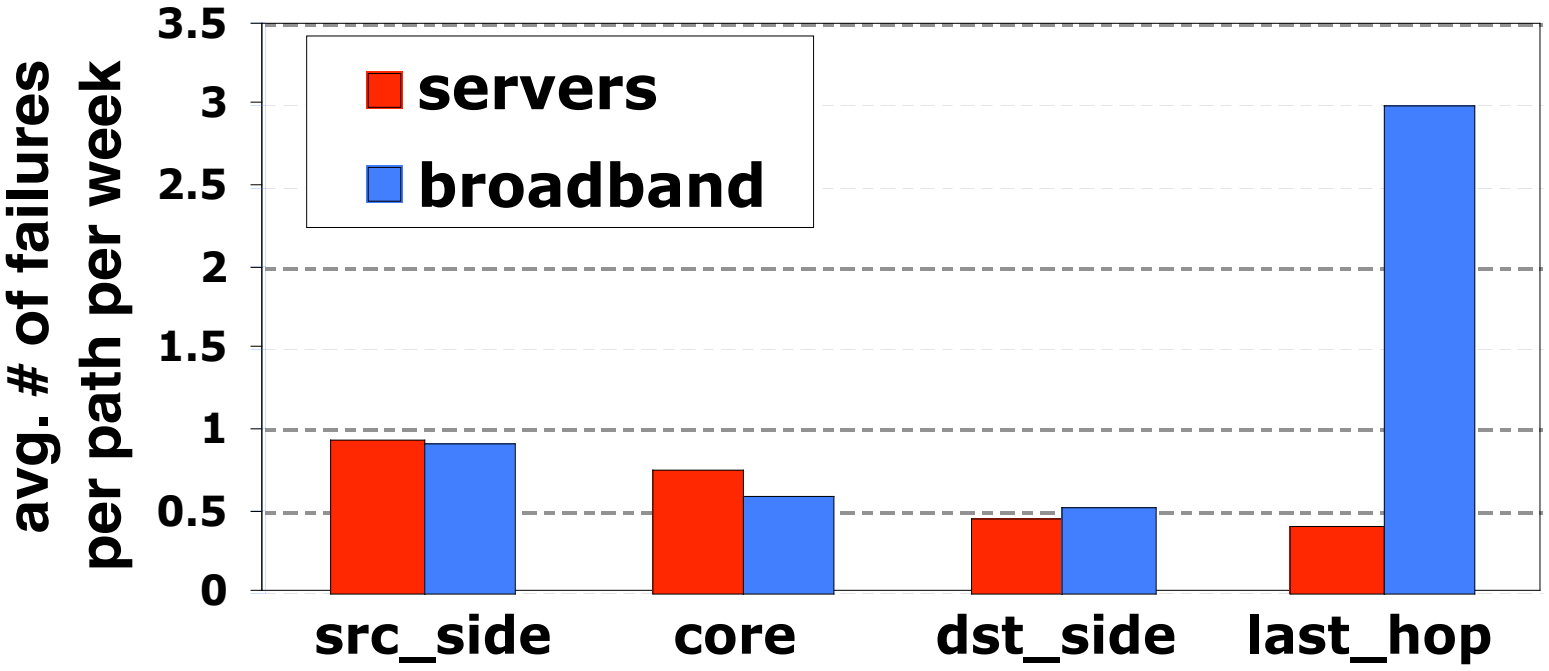
From Gummadi et al. OSDI 2004

Where do Internet paths fail?



- **Server path failures occur throughout the network**
  - very few (16%) last_hop failures

Where do Internet paths fail?



- **Most of the broadband failures happen on last_hop**
- **Excluding last_hop, server and broadband paths see similar number of failures**

From Gummadi et al. OSDI 2004

How long do Internet failures last?

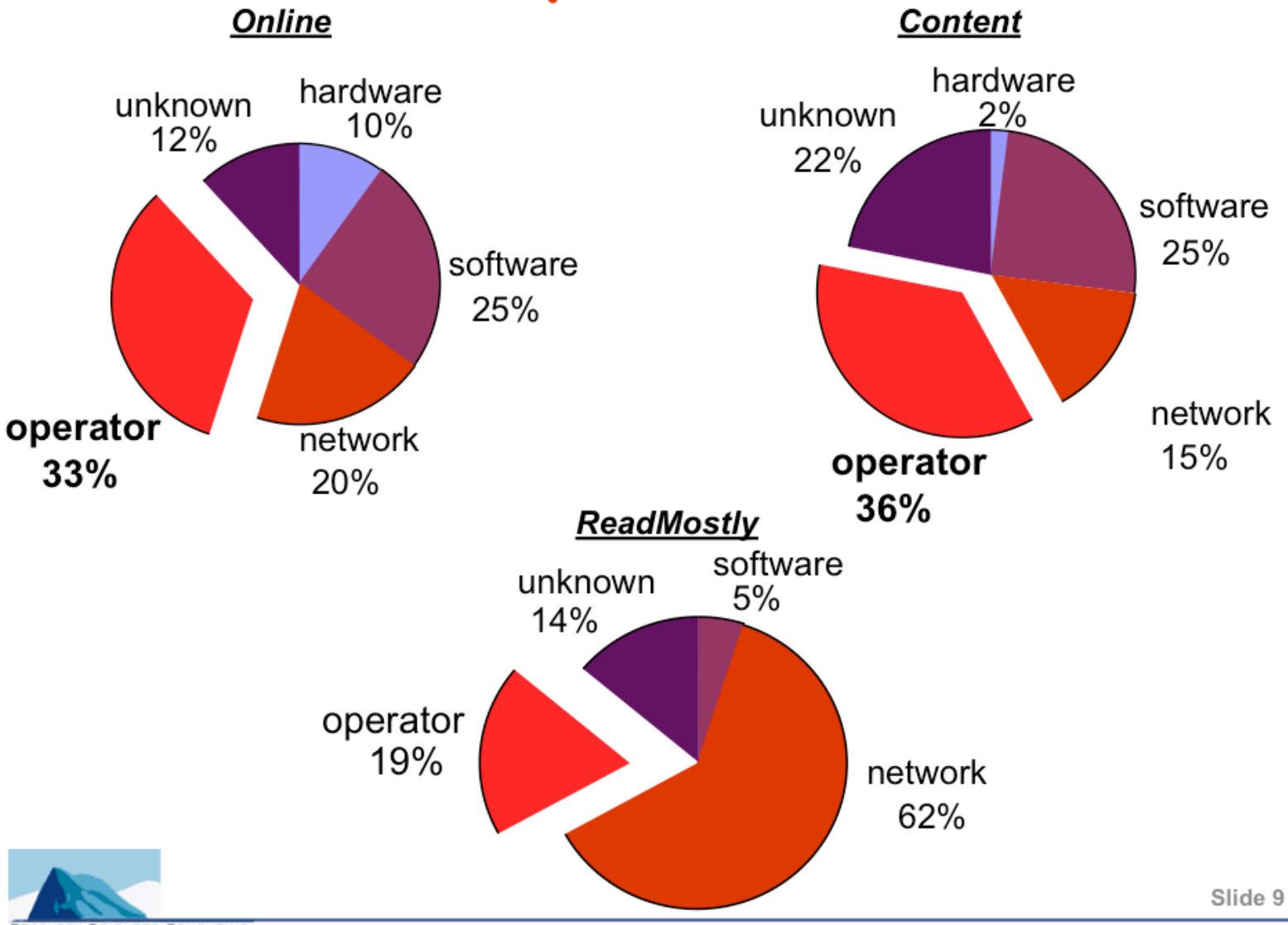
- **Failure durations are highly skewed**
- **Majority of failures are short**
  - median failure duration: 1-2 min for all paths
  - median path availability: 99.9% for all paths
- **A non-negligible fraction of paths see long failures**
  - tend to occur on last_hop
  - mean path availability: 99.6% (servers) + 94.4% (broadband)

Internet Servers Availability

Comparing the three services

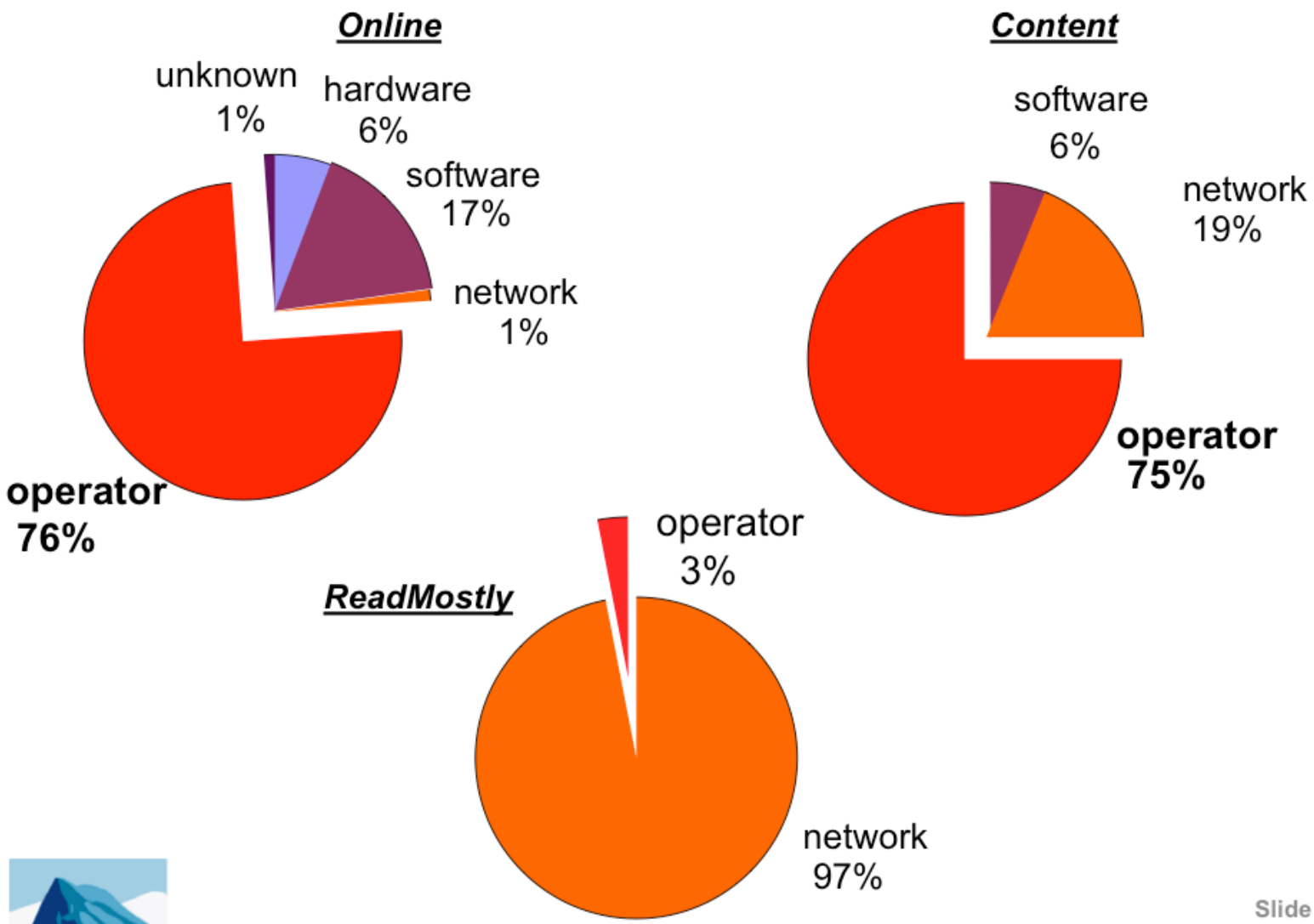
characteristic	*Online*	*ReadMostly*	*Content*
hits per day	~100 million	~100 million	~7 million
# of machines	~500 @ 2 sites	> 2000 @ 4 sites	~500 @ ~15 sites
front-end node architecture	custom s/w; Solaris on SPARC, x86	custom s/w; open-source OS on x86	custom s/w; open-source OS on x86;
back-end node architecture	Network Appliance filers	custom s/w; open-source OS on x86	custom s/w; open-source OS on x86
period studied	7 months	6 months	3 months
# component failures	296	N/A	205
# service failures	40	21	56

Failure cause by % of service failures



RECOVERY-ORIENTED COMPUTING

Failure cause by % of TTR



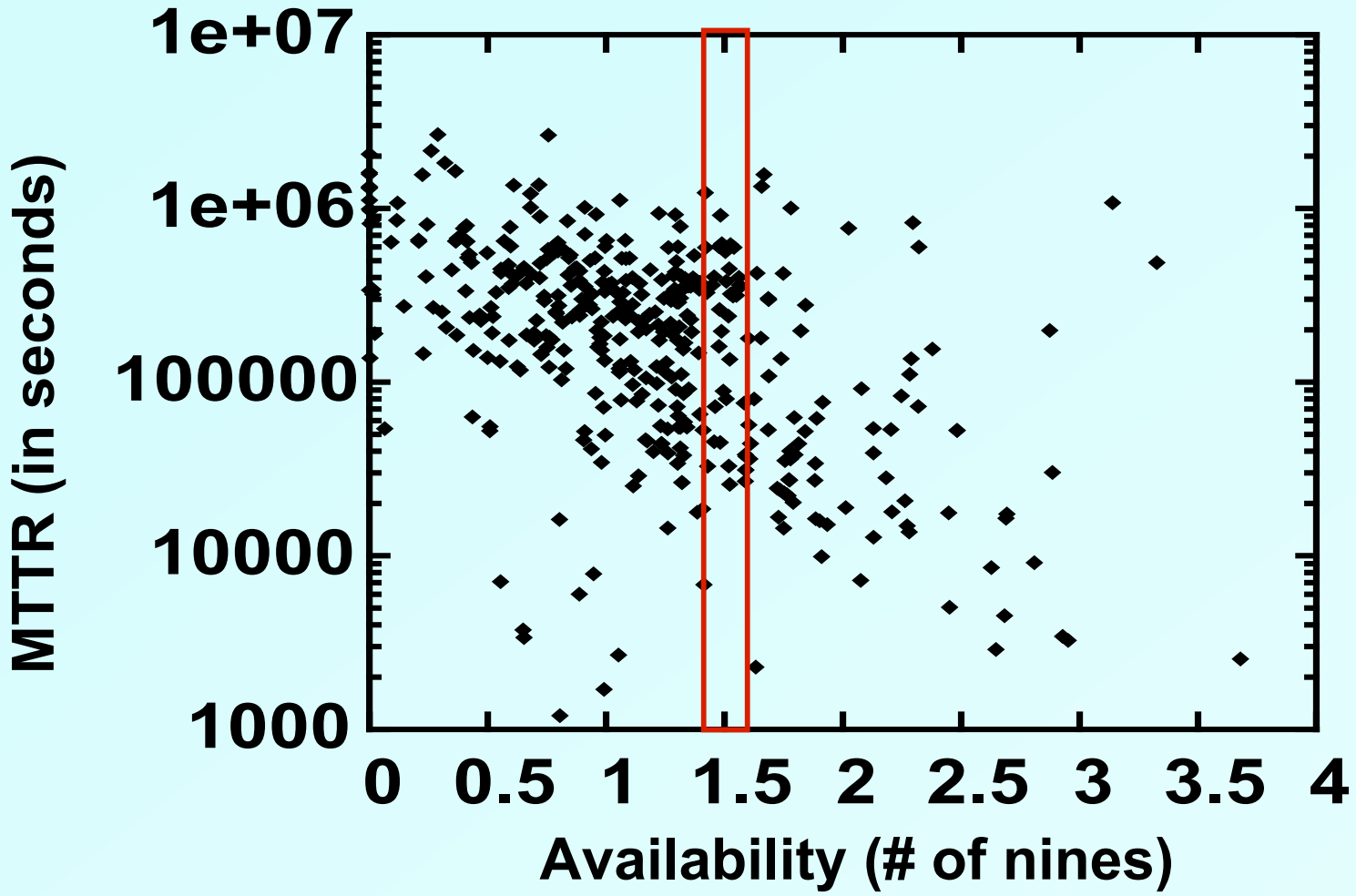
Most important failure root cause?

- **Operator error generally the largest cause of service failure**
  - Even more significant as fraction of total “downtime”
  - Configuration errors > 50% of operator errors
  - Generally happened when making changes, not repairs
- **Network problems significant cause of failures**

Wide-area systems' availability

Does Higher Availability ➔ Lower MTTR?

- **PlanetLab**



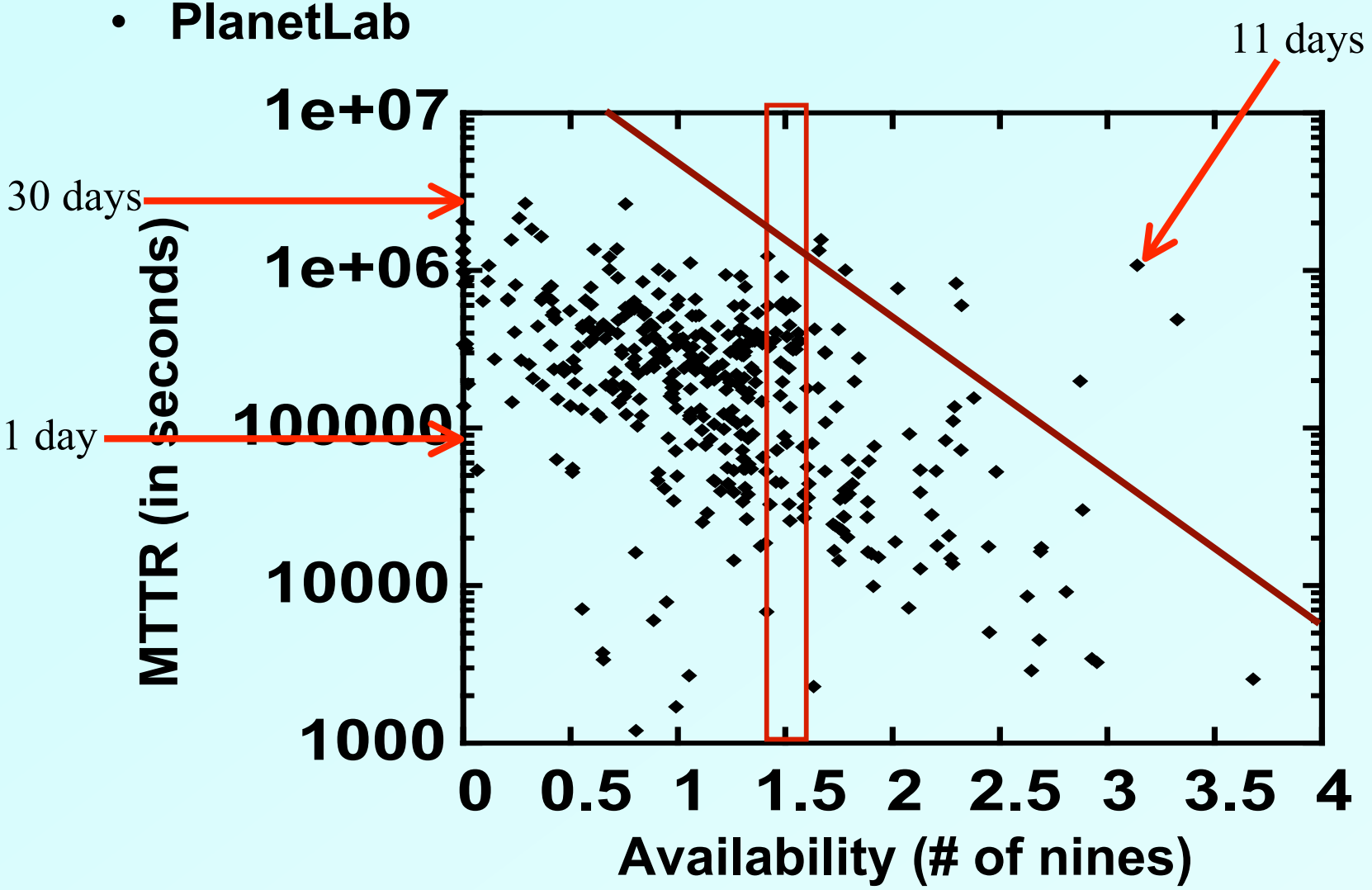
From Yalagandula et al. Worlds 2004

Result?

- **[…] there is a general trend toward better MTTR and MTTF (especially for MTTR) when availability increases.**

Does Higher Availability ➔ Lower MTTR?

- **PlanetLab**



From Yalagandula et al. Worlds 2004

Discussion

- **Operator error largest cause of service failures**
  - Is this good or bad news?

Discussion

- **Operator error largest cause of service failures**
  - Is this good or bad news?
    - Good news: software reliability is not the problem
    - Bad news: software manageability is the problem

Discussion

- **How much does it cost to add an additional 9 to a service?**

Discussion

- **Should we build fault-tolerance into our clusters:**
  - Vertically?
    - Redundant hardware
    - More sophisticated FT schemes?
  - Horizontally?
    - Wide-area distributed servers (e.g., Akamai)