Cloudlets

Enabling the Post-PC World

Mahadev Satyanarayanan
School of Computer Science
Carnegie Mellon University

Credits

Kiryong Ha, Wolfgang Richter, Yoshihisa Abe, Jan Harkes, Benjamin Gilbert
CMU-SCS: Dan Siewiorek, Martial Hebert
CMU-SEI: Grace Lewis, Soumya Simanta, Edwin Morris
Intel: Padmanabhan Pillai
Microsoft: Victor Bahl
Google: Roy Want
AT&T Research: Ramon Caceres
Lancaster University: Nigel Davies, Sarah Clinch

The world craves mobile computing!

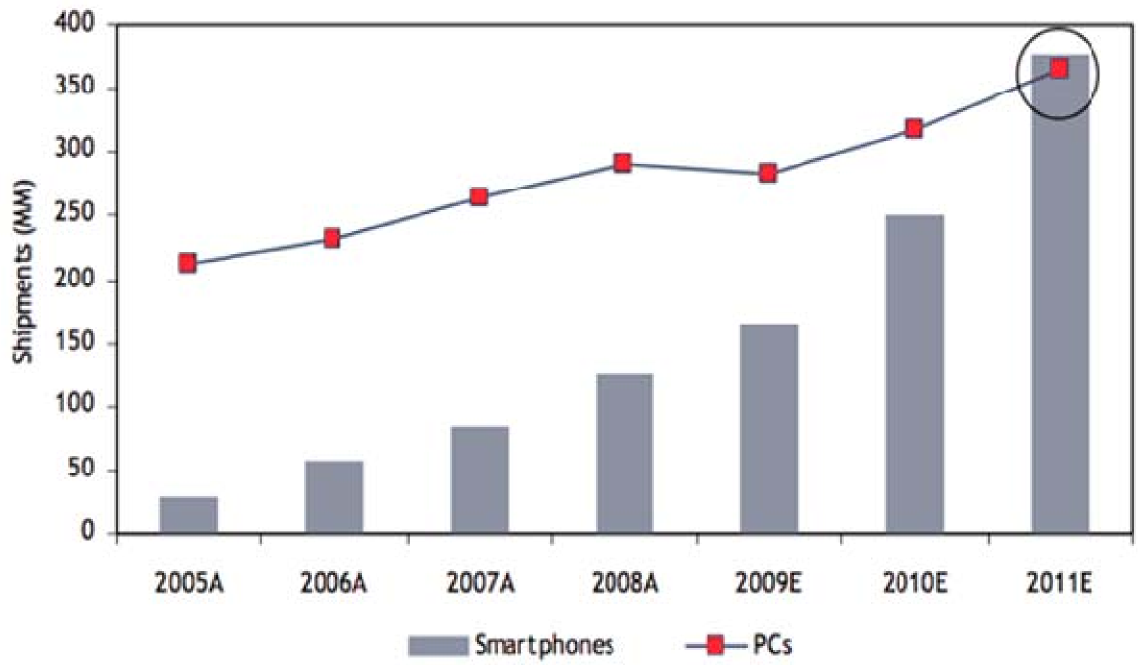
Source: eTForecasts, January 2010

U.S. and Worldwide PC Market Segments							
Unit Sales	1990	1995	2000	2005	2008	2010	2014
U.S. PC Server Sales (#M)	0.04	0.51	2.5	3.6	4.1	4.0	4.4
U.S. Desktop PC Sales (#M)	8.4	16.8	33.1	36.0	32.0	25.4	22.6
U.S. Mobile PC Sales (#M)	1.1	4.1	10.4	22.4	34.8	44.5	59.8
Worldwide PC Server Sales (#M)	0.06	0.94	5.6	10.5	14.1	14.3	17.0
Worldwide Desktop PC Sales (#M)	21.7	47.1	97.8	130.4	138.1	116.6	103.6
Worldwide Mobile PC Sales (#M)	2.4	10.0	28.5	66.3	126.0	169.8	264.0

Note: Mobile PCs include all laptop, notebook, netbook and other mobile PCs. The emerging tablet PCs and wearable PCs are also included in the mobile PC segment. PDAs and Smartphones are excluded.

Worldwide: PCs vs. Smartphones

PCs = (Desktops + Laptops + PDAs)



Source: RBC Capital Markets estimates

Chart date: August 21, 2009 *A = actual E = estimate*

How are the smartphones used?

Voice, email, chat, texting, Skype, …

Web browsing, Google/Bing search, Google maps, …

Location-based pop-ups …

Image and video capture …

???

Essentially desktop/laptop tasks on small machines!

(a bit like early TV content in the transition from radio)

Post-PC World

Mobile Cognitive Assistance

A Landmark in Computing

2011

IBM's Watson

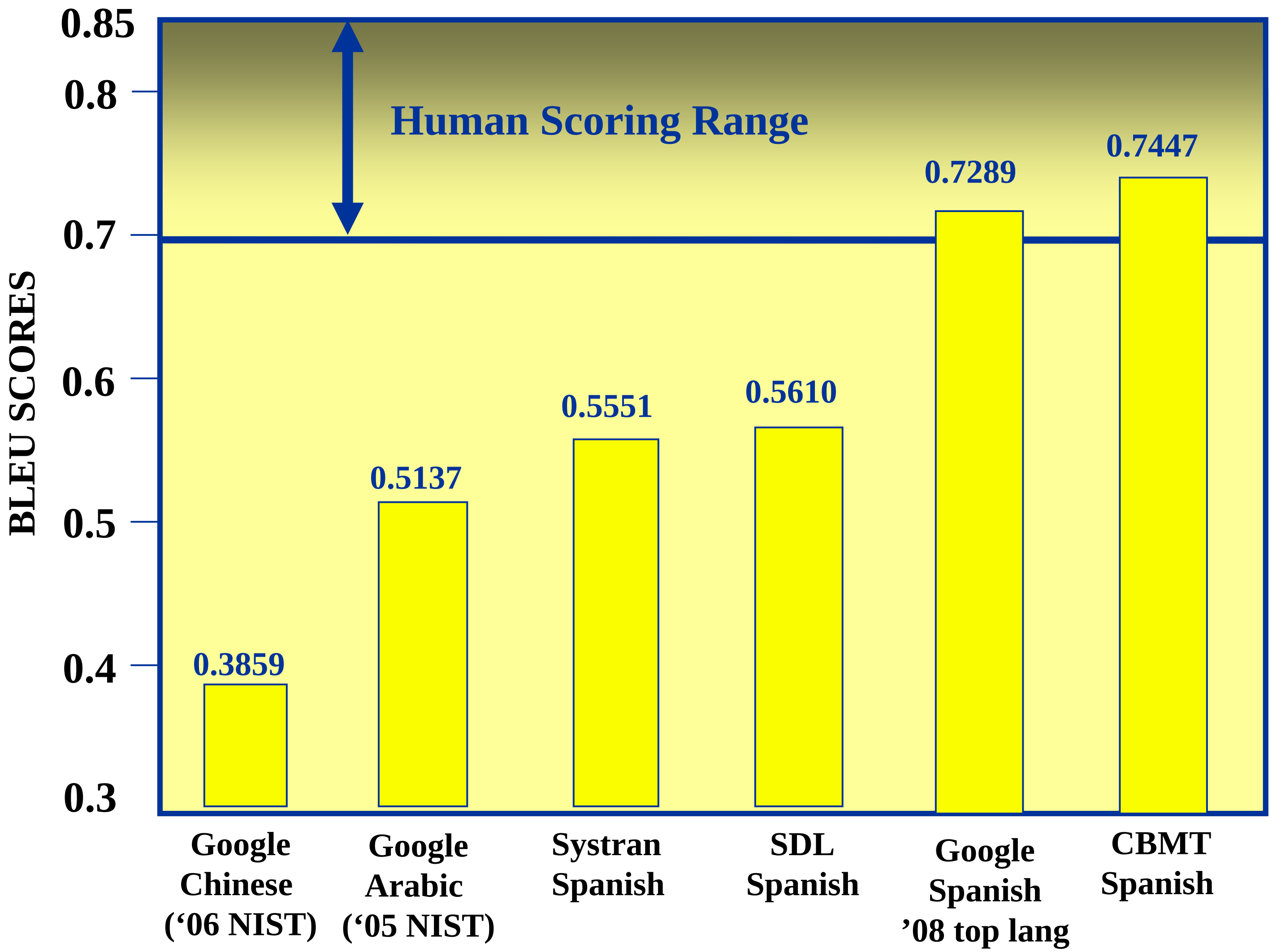
A decade earlier …

2000

IBM's Linux Wristwatch

Can I have Watson on my wristwatch?

Machine Translation Today



Based on same Spanish test set →

(slides from Carbonell, 2008)

CARBONELL, J., KLEIN, S., MILLER, D., STEINBAUM, M., GRASSIANY, T., AND FREY, J. Context-based Machine Translation. In *Proc. of the 7th Conf. of the Assoc. for Machine Translation in the Americas* (Cambridge, MA, August 2006).

Face Recognition Today

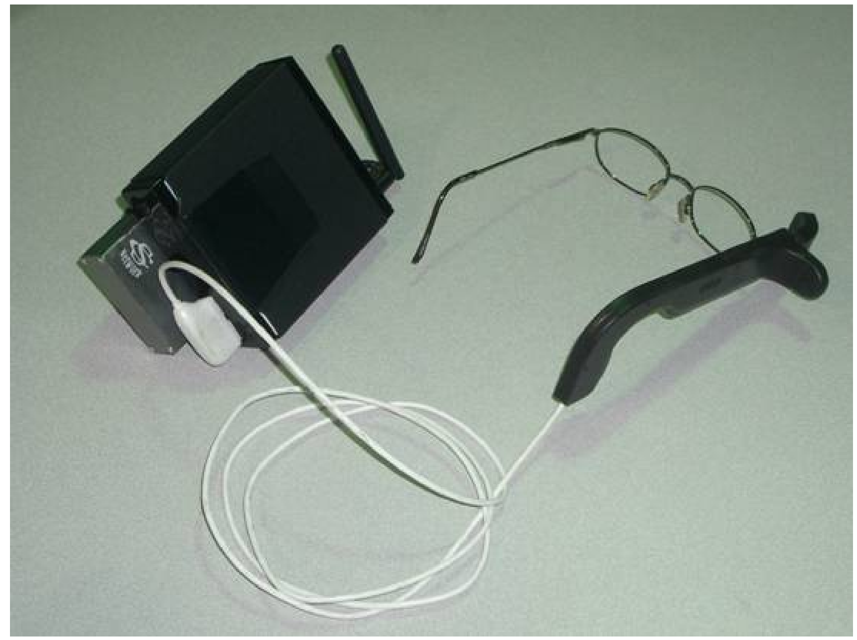
Year	Computer worse than human (%)	Computer better than human (%)	Indeterminate (%)	Worse/ Better
1999	87.5	4.2	8.3	21.0
2001	87.5	8.3	4.2	10.5
2003	45.8	16.7	37.5	2.75
2005	37.5	33.3	29.2	1.13
2006	29.2	37.5	33.3	0.78

ADLER, A., AND SCHUCKERS, M. E. Comparing Human and Automatic Face Recognition Performance. *IEEE Transactions on Systems, Man, and Cybernetics — Part B: Cybernetics 37*, 5 (October 2007).

Mobile Cognitive Assistance?

Quality of Life
Technology Center
a National Science Foundation
Engineering Research Center

Carnegie Mellon University
University of Pittsburgh



"First-Person" or "Inside-Out" vision devices that can see exactly what a person is looking at can help a robotic assistant to predict what tasks that person is trying to perform.

What’s The Catch?

These are resource-intensive tasks

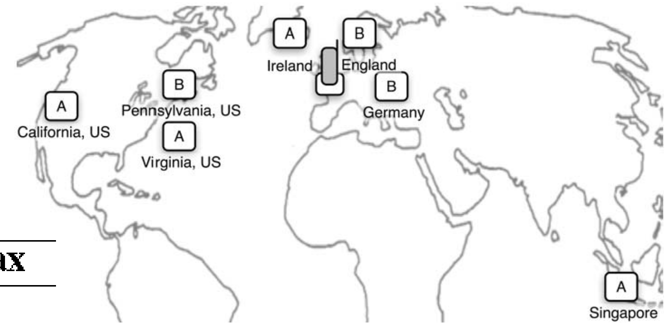
- state-of-art quality → room full of servers
- how do we achieve this “in the wild”?
  (on resource-poor, energy-limited mobile hardware)

Leverage the cloud!

But your cloud may be far away …

Amazon VNC Screen Updates

Clinch et al, 2011



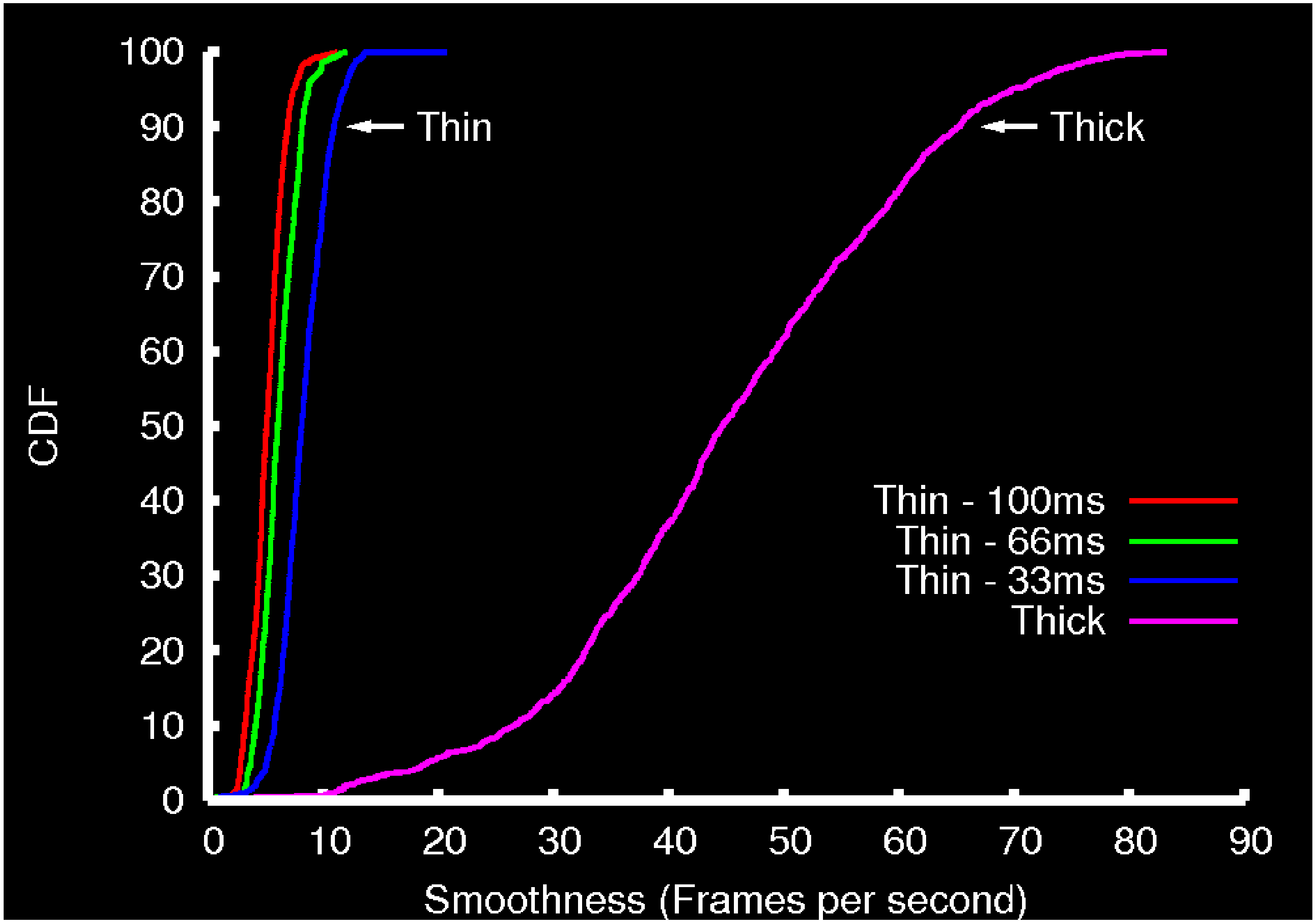
VNC client at Lancaster, UK; Wi-Fi first hop;
(key down + key up) → screen update
irl = Dublin, use = Virginia, usw = northern CA, asia = Singapore

Site	Min	1Q	Median	Mean	3Q	IQR	Max
Including TCP retries							
ec2-asia	254	271	319	692	429	158	18160
ec2-usw	212	224	227	394	235	11	12750
ec2-use	135	156	161	295	169	13	9956
ec2-irl	72	87	90	187	96	8	14710
Excluding TCP retries							
ec2-asia	254	270	284	337	422	152	488
ec2-usw	212	224	226	228	230	6	272
ec2-use	135	155	159	161	164	8	227
ec2-irl	72	87	90	91	93	6	137
Cloud processing only							
ec2-asia	55	68	71	72	74	6	98
ec2-usw	57	68	71	71	73	5	107
ec2-use	51	67	70	70	73	6	105
ec2-irl	51	66	69	68	71	5	84
Network latency (excluding retries)							
ec2-asia	191	198	212	265	354	156	402
ec2-usw	154	155	155	157	156	1	201
ec2-use	82	88	89	90	91	2	152
ec2-irl	19	20	20	22	22	2	71

Sample Internet2 RTTs (ms)

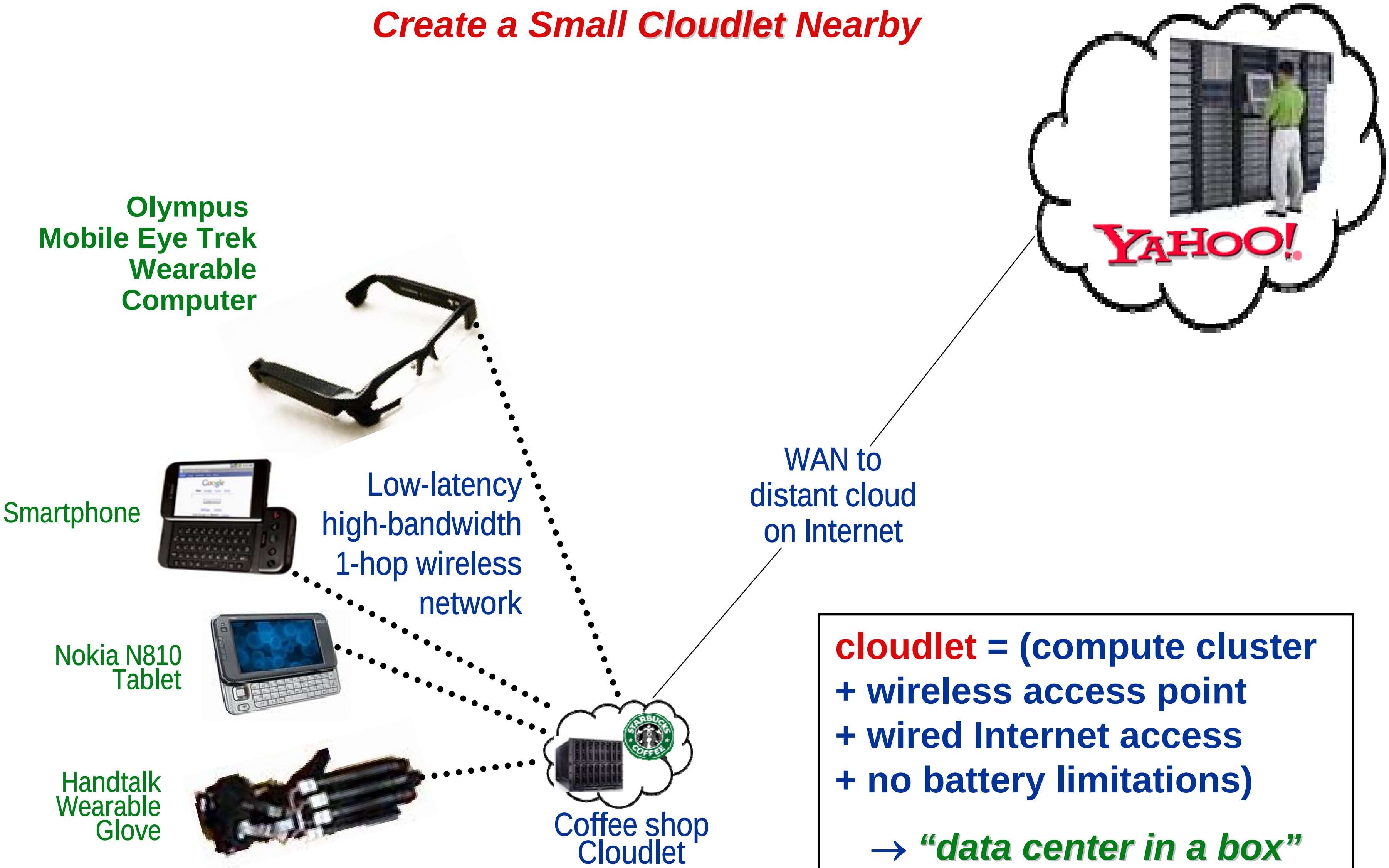
	Min	Mean	Max	Lower bound
Berkeley–Canberra	174.0	174.7	176.0	79.9
Berkeley–New York	85.0	85.0	85.0	27.4
Berkeley–Trondheim	197.0	197.0	197.0	55.6
Pittsburgh–Ottawa	44.0	44.1	62.0	4.3
Pittsburgh–Hong Kong	217.0	223.1	393.0	85.9
Pittsburgh–Dublin	115.0	115.7	116.0	42.0
Pittsburgh–Seattle	83.0	83.9	84.0	22.9

Latency Hurts (even at 100 Mbps!)



Bring the Cloud Closer!

Create a Small Cloudlet Nearby



cloudlet = (compute cluster
+ wireless access point
+ wired Internet access
+ no battery limitations)

→ *"data center in a box"*

Cloudlet vs. Cloud

	Cloudlet	Cloud
State	*Only soft state*	*Hard and soft state*
Management	Appliance model (self-managed)	Utility model (professionally administered)
Environment	"Data center in a box" (at customer premises)	Machine room (power conditioning and cooling)
Ownership	Decentralized ownership (but possibly AT&T/Verizon/…)	Centralized ownership (by Amazon, Microsoft, …)
Network	LAN latency and bandwidth	Internet latency and bandwidth
Sharing	Few users at a time	100s to 1000s of users

Key Challenge

Cloudlet provider viewpoint

- **centralization simplifies management**
- **dispersion ⇒ extreme standardization**
  minimal system management

Mobile user viewpoint

- **cloudlet software must exactly match mobile client**
- **many areas of customization even with COTS software**
  personal preferences, speech tuning, domain-specific vocabulary, …
- **“nearly right” is not good enough**

Inherent tension at global scale

Solution: “Bring Your Own VM”

Transient Customization

Delivering full VM to cloudlet: too big, too slow

Solution: assemble VM on the fly → *dynamic VM synthesis*

- cloudlet prefetches large, widely-used VM *(base VM)*
- mobile device delivers small patch just before use *(VM overlay)*
- cloudlet discards VM after use
  (or caches for future reuse)

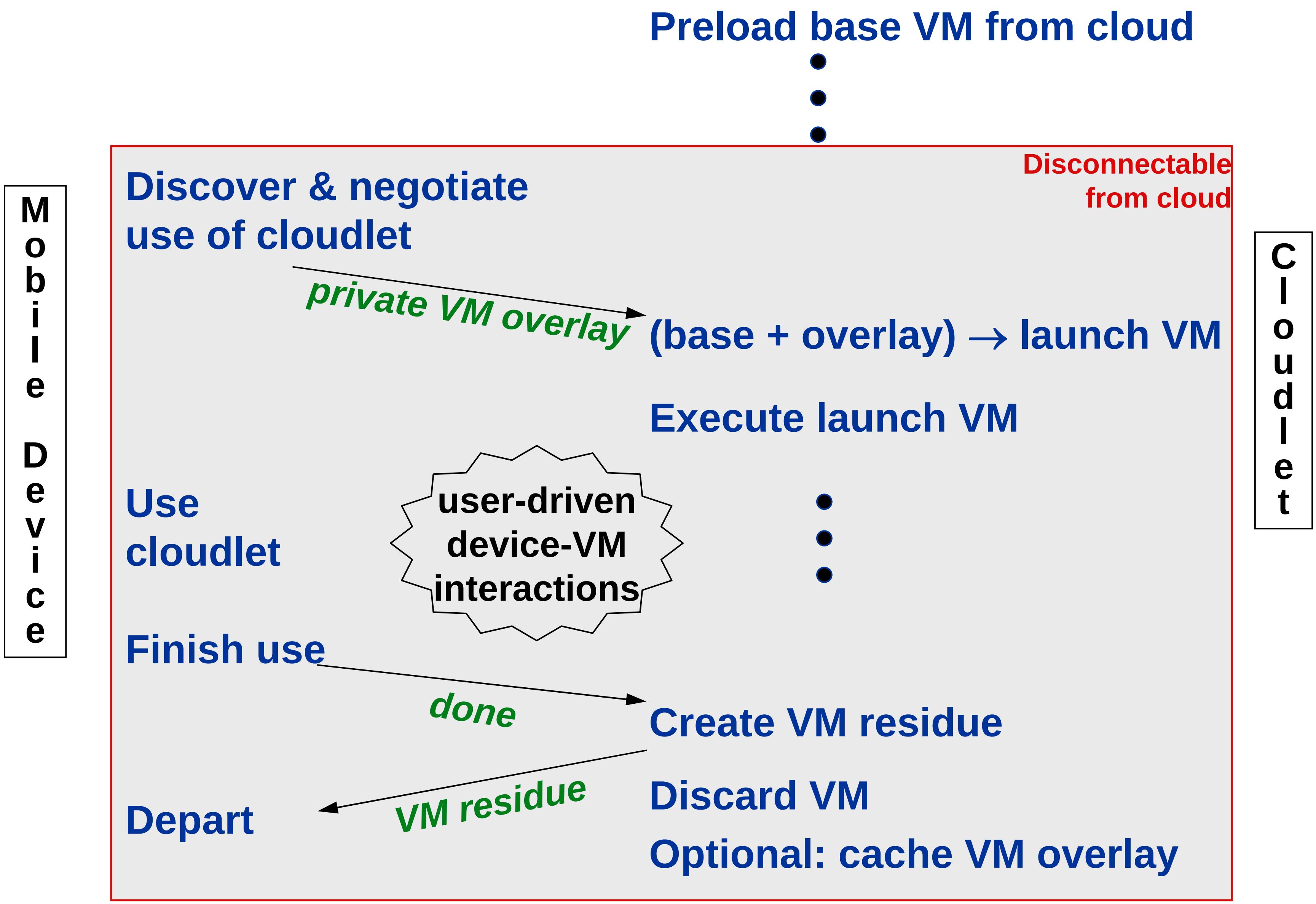
Typical overlay much smaller than base

(two orders of magnitude smaller in our experiments)

VM overlay can come from

- mobile device over wireless link, or
- from cloud over wired link (under control of mobile device)

Dynamic VM Synthesis



Other Possibilities

1. Download VM to cloudlet from cloud
   - 3GB @ 10 Mbps → ~ 2400 seconds (~40 min)

2. Upload VM from mobile to cloudlet
   - 3GB @ 100 Mbps → ~ 240 seconds (~4 min and burns battery)

3. Demand-page VM from cloud ?

4. Demand-page VM from mobile ?
   - (burns battery, but disconnectable just like VM synthesis)

Example Applications

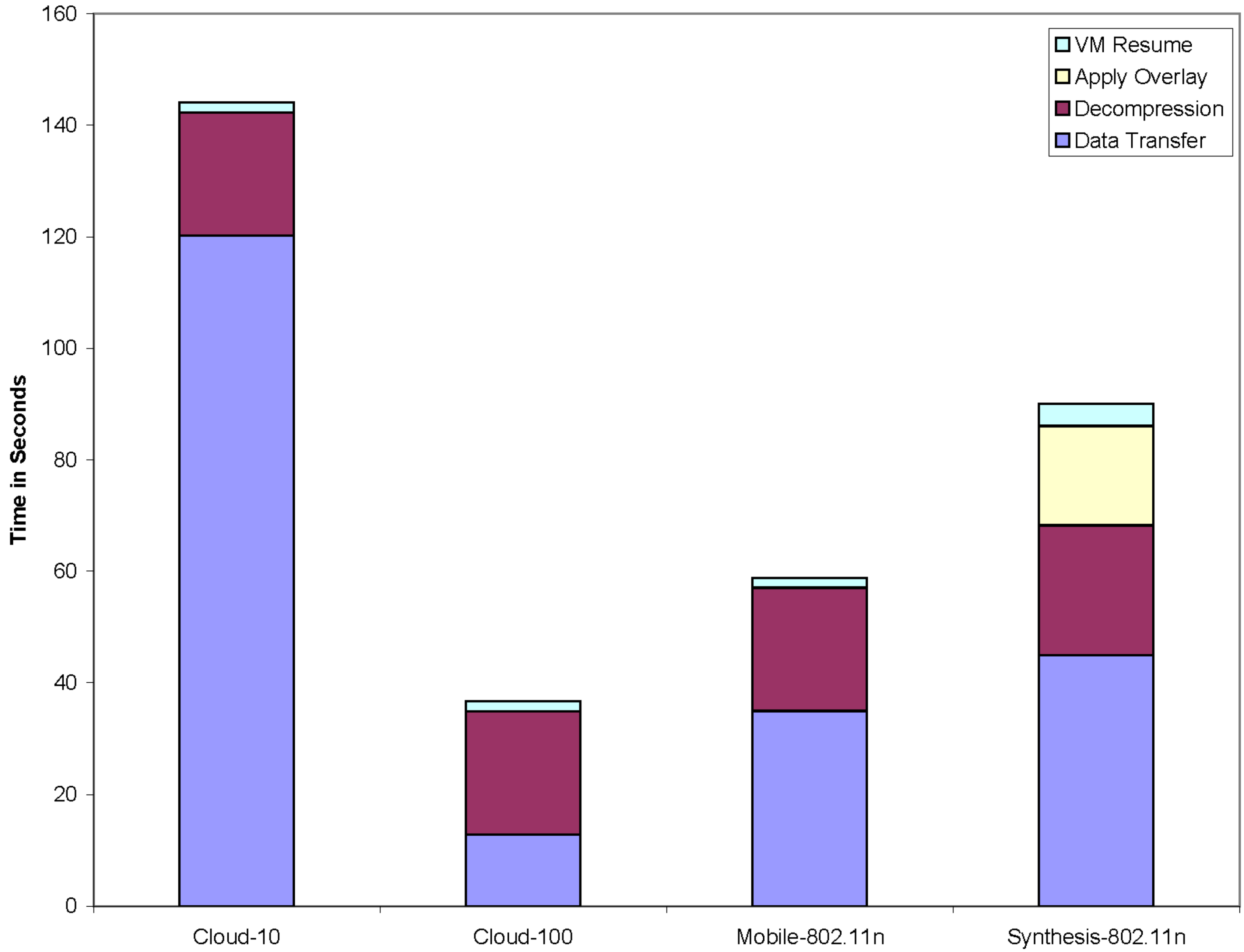
MOPED: near real-time object recognition

FACEREC: near real-time face recognition

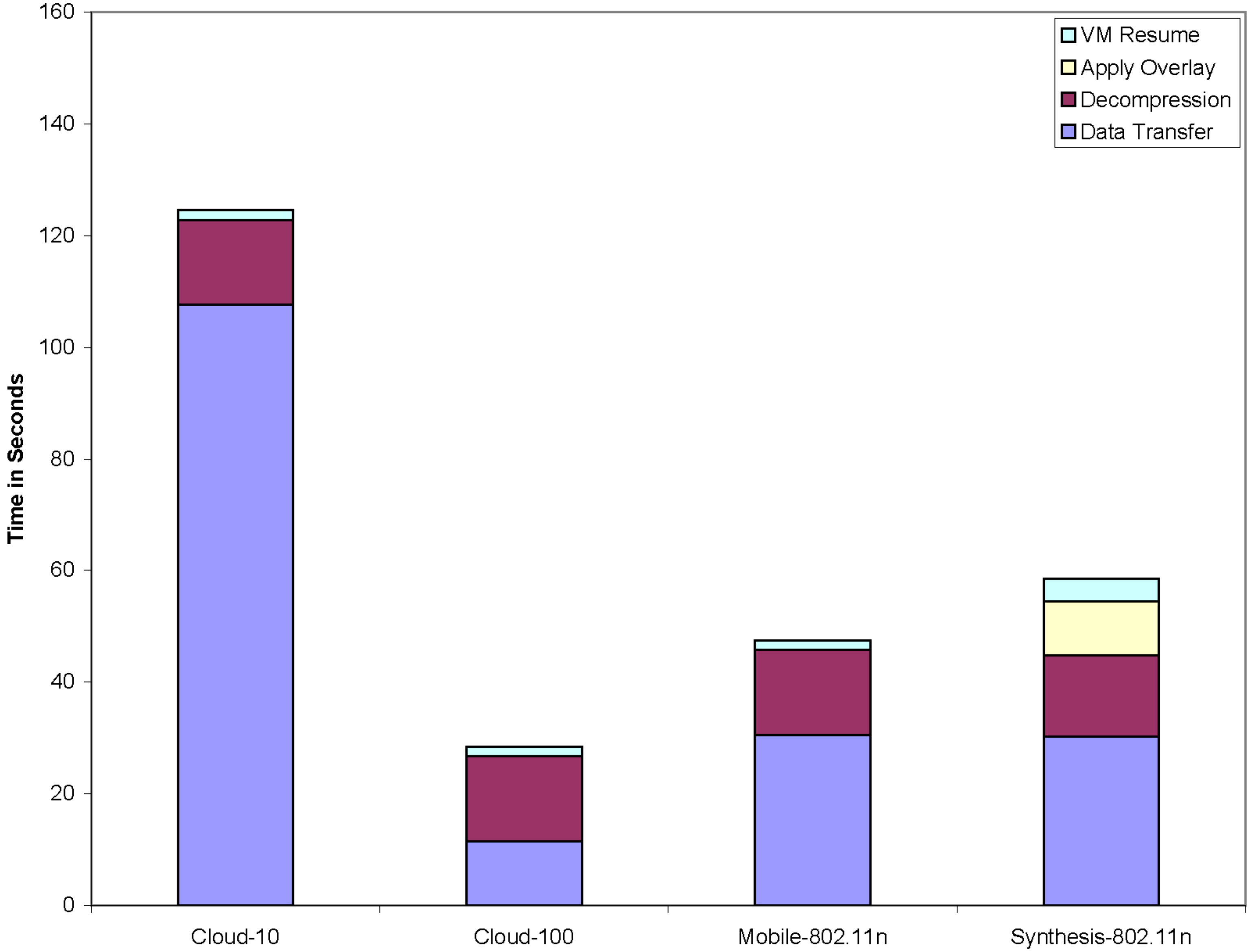
NULL: for overhead measurement

	Guest OS	Application Size (MB)	Base VM Disk (GB)	Base VM Memory (MB)	Compressed Disk Overlay (MB)	Compressed Memory Overlay (MB)
MOPED	Ubuntu 10.04	27.5 (9.7 MB for binary, 17.8 for lib)	2.5	476	27	146
FACE	Windows XP	17.66	2.1	279	58	48
NULL	Ubuntu 10.04	0	2.5	476	0.042	0.277

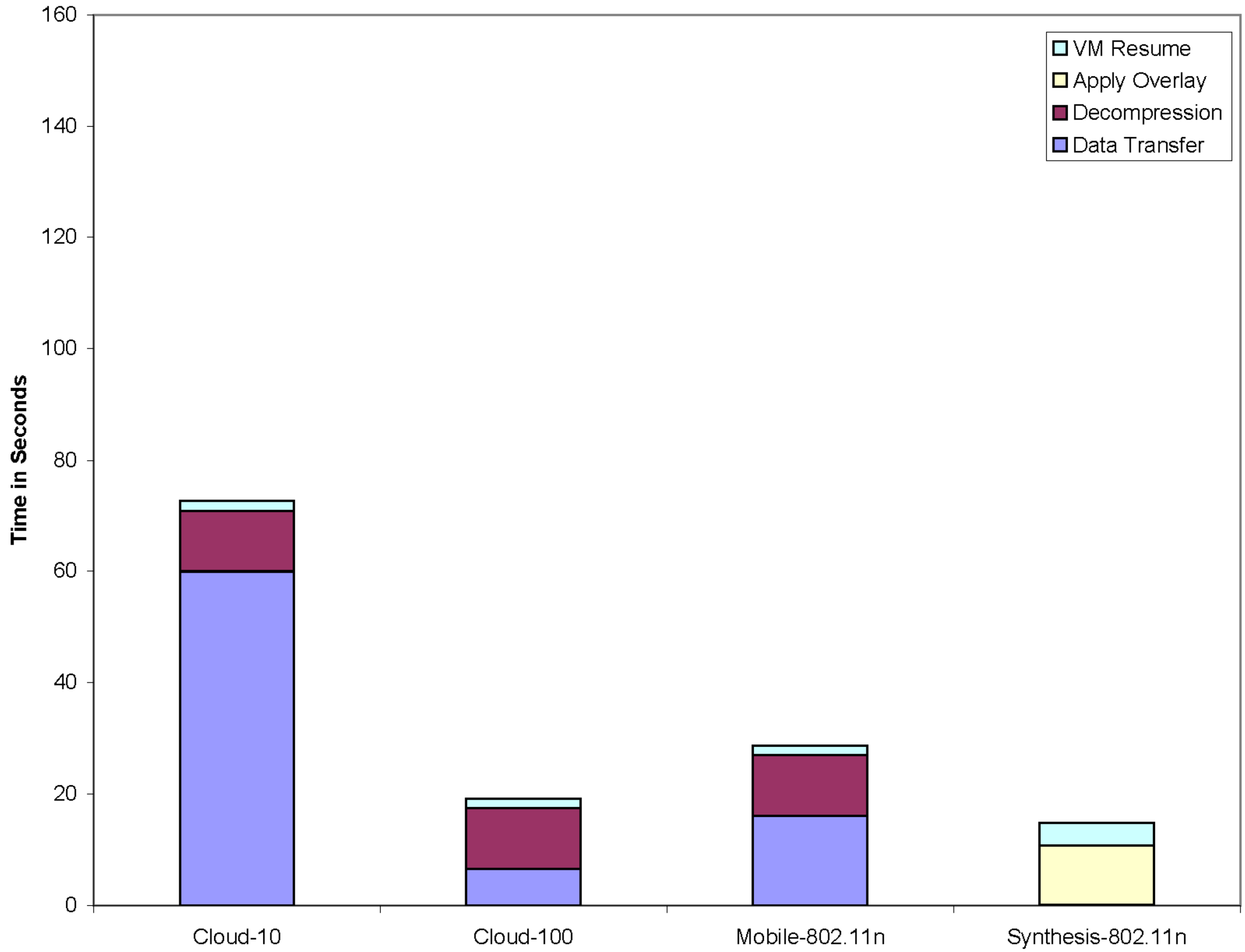
MOPED VM Launch



FACEREC VM Launch



NULL VM Launch



Some Strategic Implications

New hardware opportunities

- "data center in a box" designs
- tamper-evident cloudlets
- monitoring services

New demand for edge-only wireless bandwidth

- not limited by end-to-end bandwidth to cloud
- opportunities for new wireless technologies
  (e.g. UWB, 60 GHz radio, …)
- short range is not an issue

Cloud-cloudlet bandwidth demand

- different workload from classic edge bandwidth consumers
- high peak to average variance
- opportunities for speculation, prefetching, traffic shaping, …