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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit Sales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>U.S. PC Server Sales (#M)</td>
<td>0.04</td>
<td>0.51</td>
<td>2.5</td>
<td>3.6</td>
<td>4.1</td>
<td>4.0</td>
<td>4.4</td>
</tr>
<tr>
<td>U.S. Desktop PC Sales (#M)</td>
<td>8.4</td>
<td>16.8</td>
<td>33.1</td>
<td>36.0</td>
<td>32.0</td>
<td>25.4</td>
<td>22.6</td>
</tr>
<tr>
<td>U.S. Mobile PC Sales (#M)</td>
<td>1.1</td>
<td>4.1</td>
<td>10.4</td>
<td>22.4</td>
<td>34.8</td>
<td>44.5</td>
<td>59.8</td>
</tr>
<tr>
<td>Worldwide PC Server Sales (#M)</td>
<td>0.06</td>
<td>0.94</td>
<td>5.6</td>
<td>10.5</td>
<td>14.1</td>
<td>14.3</td>
<td>17.0</td>
</tr>
<tr>
<td>Worldwide Desktop PC Sales (#M)</td>
<td>21.7</td>
<td>47.1</td>
<td>97.8</td>
<td>130.4</td>
<td>138.1</td>
<td>116.6</td>
<td>103.6</td>
</tr>
<tr>
<td>Worldwide Mobile PC Sales (#M)</td>
<td>2.4</td>
<td>10.0</td>
<td>28.5</td>
<td>66.3</td>
<td>126.0</td>
<td>169.8</td>
<td>264.0</td>
</tr>
</tbody>
</table>

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)

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

Source: RBC Capital Markets estimates
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

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)

Human Scoring Range

## Face Recognition Today

<table>
<thead>
<tr>
<th>Year</th>
<th>Computer worse than human (%)</th>
<th>Computer better than human (%)</th>
<th>Indeterminate (%)</th>
<th>Worse/Better</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>87.5</td>
<td>4.2</td>
<td>8.3</td>
<td>21.0</td>
</tr>
<tr>
<td>2001</td>
<td>87.5</td>
<td>8.3</td>
<td>4.2</td>
<td>10.5</td>
</tr>
<tr>
<td>2003</td>
<td>45.8</td>
<td>16.7</td>
<td>37.5</td>
<td>2.75</td>
</tr>
<tr>
<td>2005</td>
<td>37.5</td>
<td>33.3</td>
<td>29.2</td>
<td>1.13</td>
</tr>
<tr>
<td>2006</td>
<td>29.2</td>
<td>37.5</td>
<td>33.3</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Mobile Cognitive Assistance?

"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

<table>
<thead>
<tr>
<th>Site</th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>Mean</th>
<th>3Q</th>
<th>IQR</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Including TCP retries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ec2-asia</td>
<td>254</td>
<td>271</td>
<td>319</td>
<td>692</td>
<td>429</td>
<td>158</td>
<td>18160</td>
</tr>
<tr>
<td>ec2-usw</td>
<td>212</td>
<td>224</td>
<td>227</td>
<td>394</td>
<td>235</td>
<td>11</td>
<td>12750</td>
</tr>
<tr>
<td>ec2-use</td>
<td>135</td>
<td>156</td>
<td>161</td>
<td>295</td>
<td>169</td>
<td>13</td>
<td>9956</td>
</tr>
<tr>
<td>ec2-irl</td>
<td>72</td>
<td>87</td>
<td>90</td>
<td>187</td>
<td>96</td>
<td>8</td>
<td>14710</td>
</tr>
<tr>
<td>Excluding TCP retries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ec2-asia</td>
<td>254</td>
<td>270</td>
<td>284</td>
<td>337</td>
<td>422</td>
<td>152</td>
<td>488</td>
</tr>
<tr>
<td>ec2-usw</td>
<td>212</td>
<td>224</td>
<td>226</td>
<td>228</td>
<td>230</td>
<td>6</td>
<td>272</td>
</tr>
<tr>
<td>ec2-use</td>
<td>135</td>
<td>155</td>
<td>159</td>
<td>161</td>
<td>164</td>
<td>8</td>
<td>227</td>
</tr>
<tr>
<td>ec2-irl</td>
<td>72</td>
<td>87</td>
<td>90</td>
<td>91</td>
<td>93</td>
<td>6</td>
<td>137</td>
</tr>
<tr>
<td>Cloud processing only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ec2-asia</td>
<td>55</td>
<td>68</td>
<td>71</td>
<td>72</td>
<td>74</td>
<td>6</td>
<td>98</td>
</tr>
<tr>
<td>ec2-usw</td>
<td>57</td>
<td>68</td>
<td>71</td>
<td>71</td>
<td>73</td>
<td>5</td>
<td>107</td>
</tr>
<tr>
<td>ec2-use</td>
<td>51</td>
<td>67</td>
<td>70</td>
<td>70</td>
<td>73</td>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>ec2-irl</td>
<td>51</td>
<td>66</td>
<td>69</td>
<td>68</td>
<td>71</td>
<td>5</td>
<td>84</td>
</tr>
<tr>
<td>Network latency (excluding retries)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ec2-asia</td>
<td>191</td>
<td>198</td>
<td>212</td>
<td>265</td>
<td>354</td>
<td>156</td>
<td>402</td>
</tr>
<tr>
<td>ec2-usw</td>
<td>154</td>
<td>155</td>
<td>155</td>
<td>157</td>
<td>156</td>
<td>1</td>
<td>201</td>
</tr>
<tr>
<td>ec2-use</td>
<td>82</td>
<td>88</td>
<td>89</td>
<td>90</td>
<td>91</td>
<td>2</td>
<td>152</td>
</tr>
<tr>
<td>ec2-irl</td>
<td>19</td>
<td>20</td>
<td>20</td>
<td>22</td>
<td>22</td>
<td>2</td>
<td>71</td>
</tr>
</tbody>
</table>
## Sample Internet2 RTTs (ms)

<table>
<thead>
<tr>
<th>Location</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
<th>Lower bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley–Canberra</td>
<td>174.0</td>
<td>174.7</td>
<td>176.0</td>
<td>79.9</td>
</tr>
<tr>
<td>Berkeley–New York</td>
<td>85.0</td>
<td>85.0</td>
<td>85.0</td>
<td>27.4</td>
</tr>
<tr>
<td>Berkeley–Trondheim</td>
<td>197.0</td>
<td>197.0</td>
<td>197.0</td>
<td>55.6</td>
</tr>
<tr>
<td>Pittsburgh–Ottawa</td>
<td>44.0</td>
<td>44.1</td>
<td>62.0</td>
<td>4.3</td>
</tr>
<tr>
<td>Pittsburgh–Hong Kong</td>
<td>217.0</td>
<td>223.1</td>
<td>393.0</td>
<td>85.9</td>
</tr>
<tr>
<td>Pittsburgh–Dublin</td>
<td>115.0</td>
<td>115.7</td>
<td>116.0</td>
<td>42.0</td>
</tr>
<tr>
<td>Pittsburgh–Seattle</td>
<td>83.0</td>
<td>83.9</td>
<td>84.0</td>
<td>22.9</td>
</tr>
</tbody>
</table>
Latency Hurts  (even at 100 Mbps!)

![Graph showing the comparison between thin and thick smoothness with latency]

- Thin - 100ms
- Thin - 66ms
- Thin - 33ms
- Thick

CDF vs Smoothness (Frames per second)
Bring the Cloud Closer!

Create a Small Cloudlet Nearby

Olympus Mobile Eye Trek
Wearable Computer

Smartphone

Low-latency high-bandwidth 1-hop wireless network

Nokia N810 Tablet

Handtalk Wearable Glove

WAN to distant cloud on Internet

Coffee shop Cloudlet

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

→ “data center in a box”
# Cloudlet vs. Cloud

<table>
<thead>
<tr>
<th></th>
<th>Cloudlet</th>
<th>Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State</strong></td>
<td><em>Only soft state</em></td>
<td><em>Hard and soft state</em></td>
</tr>
<tr>
<td><strong>Management</strong></td>
<td>Appliance model <em>(self-managed)</em></td>
<td>Utility model <em>(professionally administered)</em></td>
</tr>
<tr>
<td><strong>Environment</strong></td>
<td>“Data center in a box” <em>(at customer premises)</em></td>
<td>Machine room <em>(power conditioning and cooling)</em></td>
</tr>
<tr>
<td><strong>Ownership</strong></td>
<td>Decentralized ownership <em>(but possibly AT&amp;T/Verizon/…)</em></td>
<td>Centralized ownership <em>(by Amazon, Microsoft, …)</em></td>
</tr>
<tr>
<td><strong>Network</strong></td>
<td>LAN latency and bandwidth</td>
<td>Internet latency and bandwidth</td>
</tr>
<tr>
<td><strong>Sharing</strong></td>
<td>Few users at a time</td>
<td>100s to 1000s of users</td>
</tr>
</tbody>
</table>
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

Preload base VM from cloud

Discover & negotiate use of cloudlet

Discover & negotiate use of cloudlet

(private VM overlay)

(base + overlay) → launch VM

Execute launch VM

Use cloudlet

user-driven device-VM interactions

Use cloudlet

Finish use

Create VM residue

Finish use

Discard VM

Depart

Optional: cache VM overlay

Depart

VM residue

Depart

VM residue

Preload base VM from cloud

Disconnectable from cloud
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

<table>
<thead>
<tr>
<th>Guest OS</th>
<th>Application Size (MB)</th>
<th>Base VM Disk (GB)</th>
<th>Base VM Memory (MB)</th>
<th>Compressed Disk Overlay (MB)</th>
<th>Compressed Memory Overlay (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOPED</td>
<td>Ubuntu 10.04 (9.7 MB for binary, 17.8 MB for lib)</td>
<td>2.5</td>
<td>476</td>
<td>27</td>
<td>146</td>
</tr>
<tr>
<td>FACE</td>
<td>Windows XP</td>
<td>2.1</td>
<td>279</td>
<td>58</td>
<td>48</td>
</tr>
<tr>
<td>NULL</td>
<td>Ubuntu 10.04</td>
<td>0</td>
<td>476</td>
<td>0.042</td>
<td>0.277</td>
</tr>
</tbody>
</table>
FACEREC VM Launch

![Bar chart showing time in seconds for different VM launches]

- **Cloud-10**: High time due to VM Resume, Apply Overlay, Decompression, and Data Transfer.
- **Cloud-100**: Moderate time due to fewer processes.
- **Mobile-802.11n**: Lower time as compared to others, with medium Decompression and Data Transfer.
- **Synthesis-802.11n**: Higher time than Mobile-802.11n due to similar processes as Cloud-10.
NULL VM Launch

Time in Seconds

- Cloud-10
- Cloud-100
- Mobile-802.11n
- Synthesis-802.11n

Legend:
- VM Resume
- Apply Overlay
- Decompression
- Data Transfer
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, …