Volunteer Computing in the Earth Sciences

Carl Christensen carlc@atm.ox.ac.uk
Chief Software Architect, Quake-Catcher Network
School of Earth Sciences, Stanford University

Senior Consultant, climateprediction.net
Department of Physics, Oxford University

STANFORD UNIVERSITY

UNIVERSITY OF OXFORD
DOMI NVS ILLV MINA TIO MEA

What is Volunteer Computing?

- A specialized form of “distributed computing” which is really an “old idea” in computer science -- using remote computers to perform a same or similar tasks
- Resources volunteered by participants from the public & some corporate users
- Was around before '99 but really took off then with SETI@home (5 million participants)
- Most projects today use the BOINC platform

What is BOINC!

- BOINC = (Cal-)Berkeley Open Infrastructure for Network Computing – offshoot of David Anderson's work on SETI@home
- Large Open Source project but the bulk of the work is done by a few core staff at Berkeley. Projects such as CPDN contributes (graphics libs, networking/libcurl layer, testing, trickles)
- SETI@Home was the pioneer and "killer" VC app with user & CPU-year numbers that are the envy of every other DC app, 5 million total users, 500K at once
- BOINC is based on the experiences of the SETI@Home team in handling millions of users, downloads and uploads, investment of >US$2million

What is BOINC! (cont'd)

- So it makes sense to use BOINC which has a "tried and tested" framework instead of keep playing "catch-up" and "reinventing the wheel"
- BOINC allows us to focus on what we do best (or should be doing best): Climate science, climate modelling, visualisation packages, cross-platform porting of models, grid applications to "clamp on" the BOINC server-side
- Reality: Everybody wants "old SETI #'s" of millions of users – but tens of thousands more likely (but "nothing to sneeze at" -- it's still "Top Supercomputer" levels of processing power/TeraFLOPS!)
- Visit http://boinc.berkeley.edu

Volunteer Computing Stats

- SETI@home peak cap with a million+ concurrent users about 1 PF = 1000 TF (years ago)
- All BOINC projects currently about 2.5PF, broke petaflop barrier 01/08
- folding@home (non-BOINC) avgs 4-5PF thanks to PS3 and GPU work
- for comparison top supercomputer (June '09 – top500.org) DOE/LLNL Roadrunner broke petaflop barrier 5/08= 1.1PF max, $120 million, 2.5MW power req
- climate*prediction*.net (CPDN) running at about 80 TF conservative estimate (40K participating 2GF machines) – about 46th on top 500 supercomputer list
- Offers high CPU power at low cost (need a few developers/sysadmins to run the "supercomputer")
- About 50 active BOINC projects, five have over 200K user, most much smaller

Volunteer Computing Challenges

- No control over quality or amount of resources
- Great for brute-force or 'embarassingly parallel' – but can't access entire CPUs & memory parallel a la supercomputer (i.e not appropriate for all problems)
- Can be risky as far as numbers needed (i.e. "if you build it, they may not come!") – may need to look into PR/media; address user attrition issues etc
- Interaction with participants can scare off more socially awkward academics and staff (successful projects have "hams" :-) – the "social aspects" take time & resources

Volunteer Computing Benefits

- Enormous potential for “free” resources from computing enthusiasts and corporate partners (at the “top supercomputer” levels and above, esp with GPU applications e.g. NVidia CUDA)
- Great for public outreach, education, publicity (QCN & CPDN in schools, BBC programmes etc)
- Relatively low cost as compared with paying for supercomputer resources (3-5 computing staff); fraction of a percent compared to supercomputer time
- Can be applied to a wide range of platforms (gaming consoles, graphics cards, all CPU O/S’, 32 & 64-bit) and problems

Case Study 1 – climate*prediction*.net

- ~10 years ago (October '99) – Myles Allen article in Nature ("Do It Yourself Climate Prediction") pondering the feasibility of large ensemble climate modelling & the SETI@home paradigm (what is now called "volunteer computing")
- 6 years ago (12th September '03) – CPDN "Classic" launched at the Science Museum in London
- ~5 years ago (August '04) – CPDN merged with the BOINC project and launched our first app…

climateprediction.net Challenges

- Climate models (ESM's, AOGCM's etc) are very large, complex systems developed by physicists sometimes over decades (& proprietary in case of UKMO)
- ~1 million lines of Fortran code (HadCM3 - 550 files, 40MB text source code)
- Little documentation (the science is often well documented but not the software and design of the system per se)

climate*prediction*.net Challenges (cont'd)

- Also "utility" code written by various scientists & students over the years (outside of model code, 220 files, 12MB source, 250K lines); often workable but hard to implement on a cross-platform PC project
- Meant to be run on supercomputers, primarily 64-bit – not designed (or indeed envisioned) to be run on anything other than a supercomputer or at the very least, a Linux cluster

climateprediction.net Overview

- Proprietary, licenced by UK MetOffice, distribute executable/binary form only
- Resolution originally used: 2.75x3.75 degrees (73 lat x 96 long), now N144 (HadAM3) also (about 1 degree square)
- Typically run on a supercomputer (i.e. Cray T3E) or 8-node Linux cluster (minimum)
- Ported to a single-processor, 32-bit Linux box, Original: Windows only, now also Mac OS X, Linux; also doing 64-bit (HadGAM) now (25% of all active users, about 13K machines, are 64-bit)
- Intel Fortran Win & Linux, IBM XLF for Mac, now all Intel compilers (Mac, Linux, Win), also looking into GNU Fortran (new versions can compile the UM)
- As of today about 70 million model-years completed (since 9/2003)

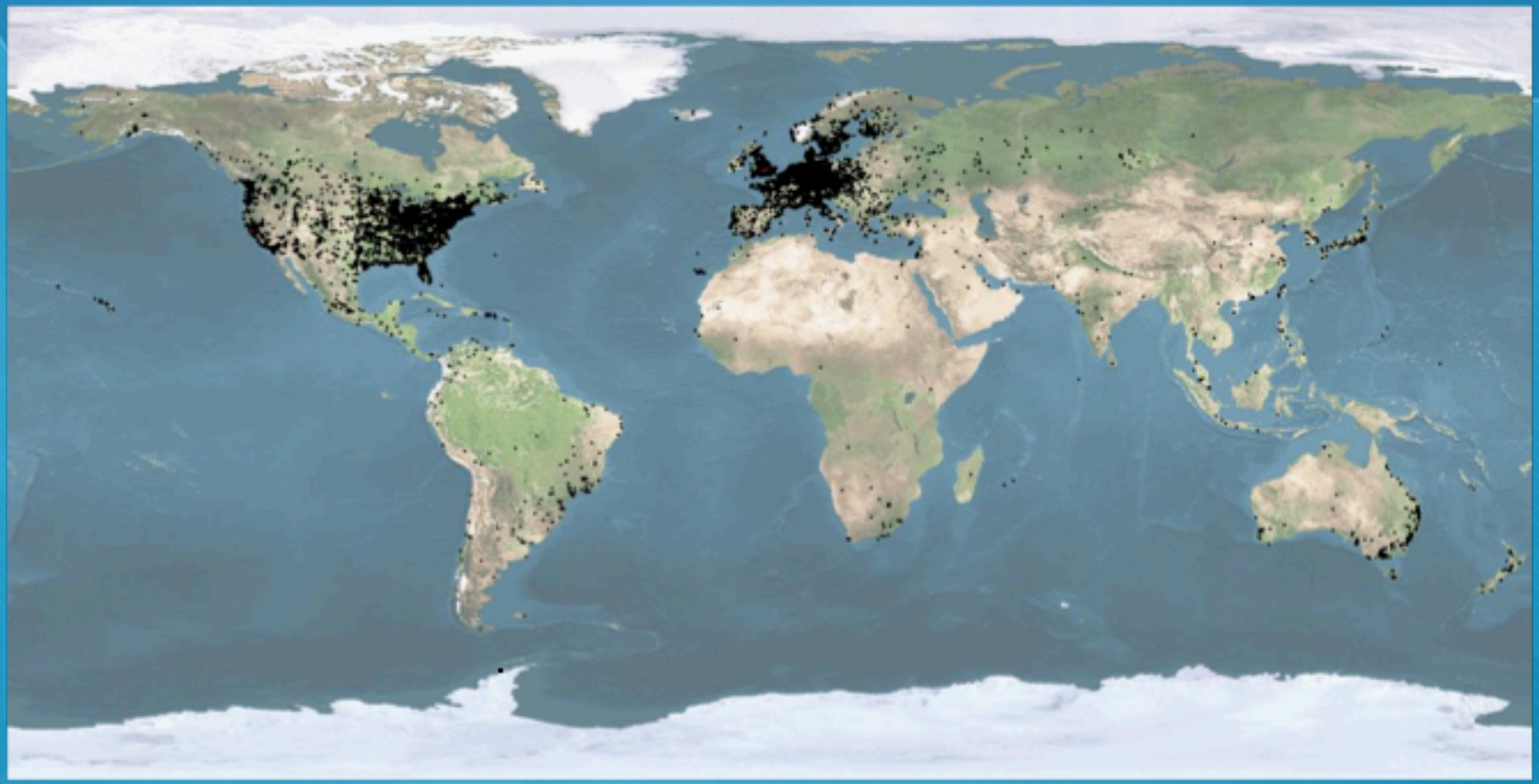
climate*prediction*.net

BBC Climate Change Experiment

- 160 year coupled-model (HadCM3) run
- Promoted as part of the BBC "Climate Chaos" season of programmes & documentaries in '06
- "Meltdown" documentary featured CPDN and launched the experiment in February of '06
- Shown on BBC4 , had two million viewers, so was also shown on BBC 1 (promos during "Eastenders"!)
- Results show featuring David Attenborough broadcast in January of '07
- Nominated for a BAFTA in "computer interactive" category (lost to Terry Pratchett's Hogfather)
- http://www.bbc.co.uk/sn/climateexperiment/

climate*prediction*.net Users Worldwide

>300,000 users total



hadcm3l version 5.15 [workunit: hadcm3ln_2bjz_00447461]

BBC

Climate Change Experiment

Use keyboard keys to change view

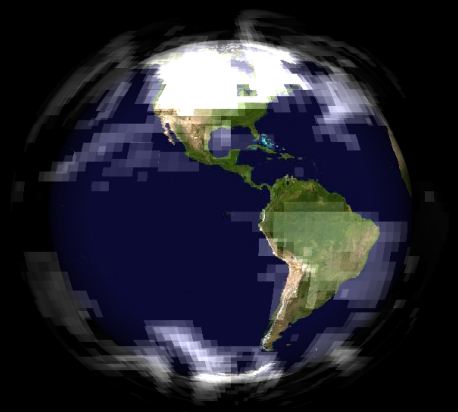
Use CTRL + key when in screensaver mode
T - Temperature
R - Rain & snow
P - Pressure
C - Clouds
S - Stop/Start rotation
G - Show/Hide grid
H - Help & more options

Current view:
Clouds & surface

This computer model of Planet Earth simulates the atmosphere & ocean on a 3-D global grid. Gridded view shows the model grid scale. Switch between cloud (C) and rain (R) views to observe your model's weather.

This globe shows your climate model running

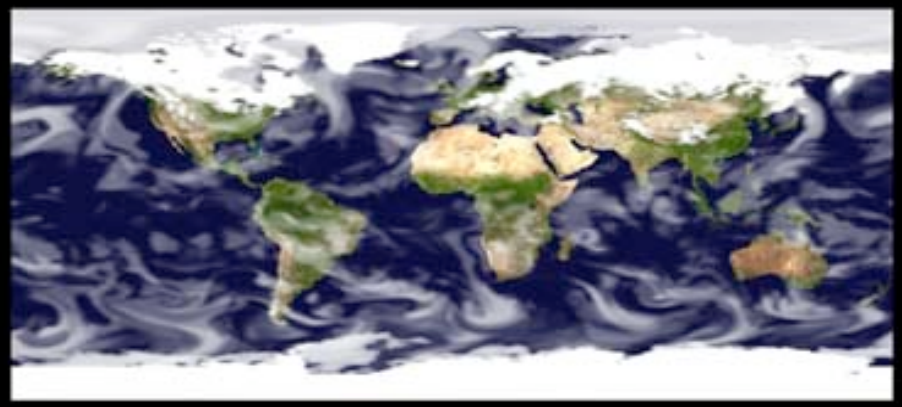
Model date and time: 26/01/1921 14:00



Thanks for taking part! Modelling the first few years is extremely useful for us, so do ke

bbc.co.uk/climatechange

created by climate*prediction*.net



hadcm3l Run ID : a2000_mth201_025a4_000
User : tolig Team : (None)
Phase : 1 of 1 / Timestep : 36602 of 51994
Model Date : 15/12/2000 04:00
CPU Time: 0304:46:48 (29.98 s/TS)
Toggle : 1=Snow, 2=LowCld, 3=MidCld, 4=HiCld

hadcm3l version 5.15 [workunit: hadcm3ln_2bjz_00447461]

BBC

Climate Change Experiment

Use keyboard keys to change view

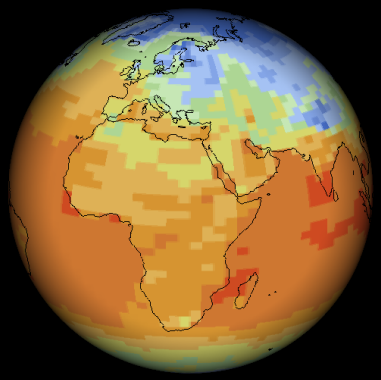
Use CTRL + key when in screensaver mode
T - Temperature
R - Rain & snow
P - Pressure
C - Clouds
S - Stop/Start rotation
G - Show/Hide grid
H - Help & more options

This computer is contributing to the biggest climate change experiment in the world

To find out more about how you are helping climate research, visit the project website:

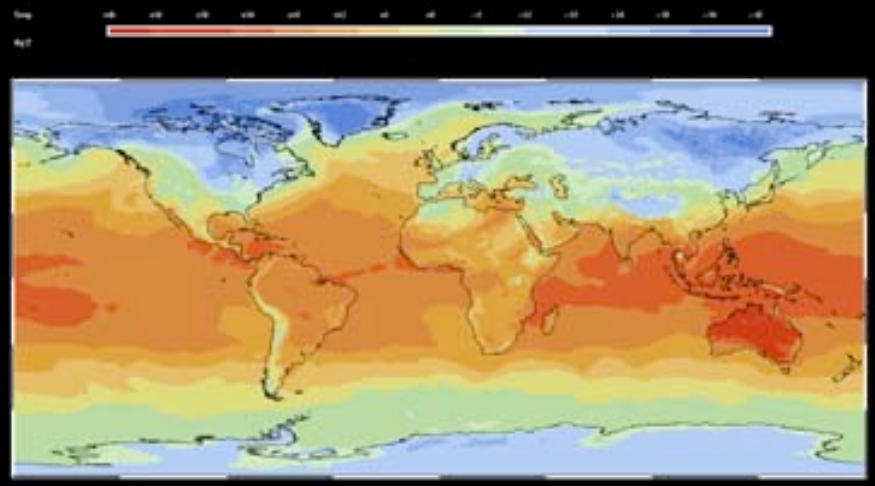
This globe shows your climate model running

Model date and time: 27/01/1921 04:00



bbc.co.uk/climatechange

created by climate*prediction*.net



hadcm3l Run ID : a2000_mth201_025a4_000
User : tolig Team : (None)
Phase : 1 of 1 / Timestep : 36600 of 51994
Model Date : 15/12/2000 04:00
CPU Time: 0304:46:03 (29.98 s/TS)
R=Rainfall, P=Pressure, S=SnowAlbCld

climate*prediction*.net
for Educational Outreach

Students at Gosford Hill School, Oxon, viewing their CPDN model

- CPDN has public education via the website, media, and schools as an important facet of the project
- Website has much information on climate change and related topics to the CPDN program.
- Open University (UK) offered a short course (S199) utilizing the climateprediction.net experiment (MS Windows client)
- Students hosted a debate on climate change issues, compared and contrasted their results, etc.

climate*prediction*.net Summary

- A good example of a high-CPU requirement volunteer computing project
- Ongoing: new experiments with MetOffice Hadley Centre models (slab model, coupled model, atmos only, regional models, newer versions i.e. HadGam/HadGem), 64-bit
- Interested in collaboration with other modelling groups to get other models on CPDN (ECHAM5?, CCSM?)
- Thanks to Myles Allen, Tolu Aina, Milo Thurston, Hiro & Kuniko Yamazaki, Sue Rosier & David Frame

STANFORD UNIVERSITY

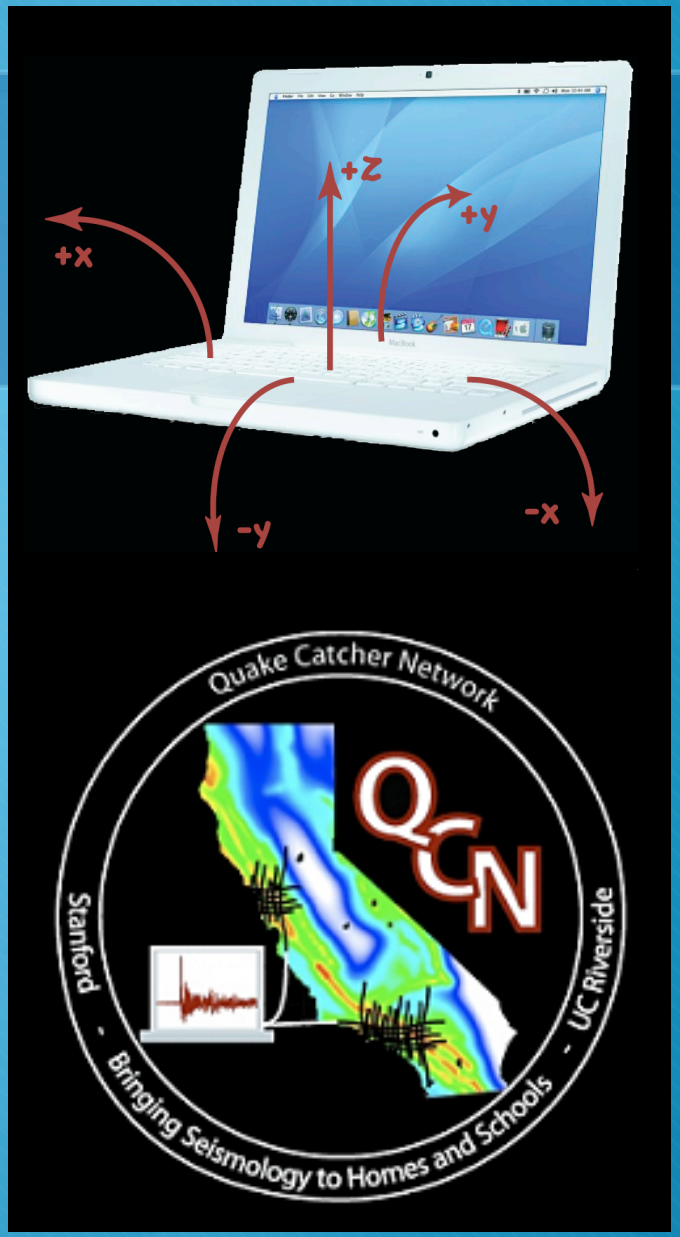
Case Study 2- The Quake-Catcher Network

Prof. Jesse Lawrence (Stanford); Prof. Elizabeth Cochran (UC-Riverside)

- Based at Stanford with collaborators in UC-Riverside and international collaboration starting with Chile
- "Opposite" the usual volunteer computing projects with high computing requirements
- Sensors report seismic events ("triggers") over the Internet to our servers via internal (laptop) or external (USB) sensors

Quake-Catcher Network

http://qcn.stanford.edu



- The Goal: To network computers with internal or USB-connected accelerometers for rapid earthquake detection.
- The Method: We use volunteer computing to monitor sensors internal or connected to computers when they are not otherwise being used.
- Potential for low-cost early warning systems; also great for educational initiatives

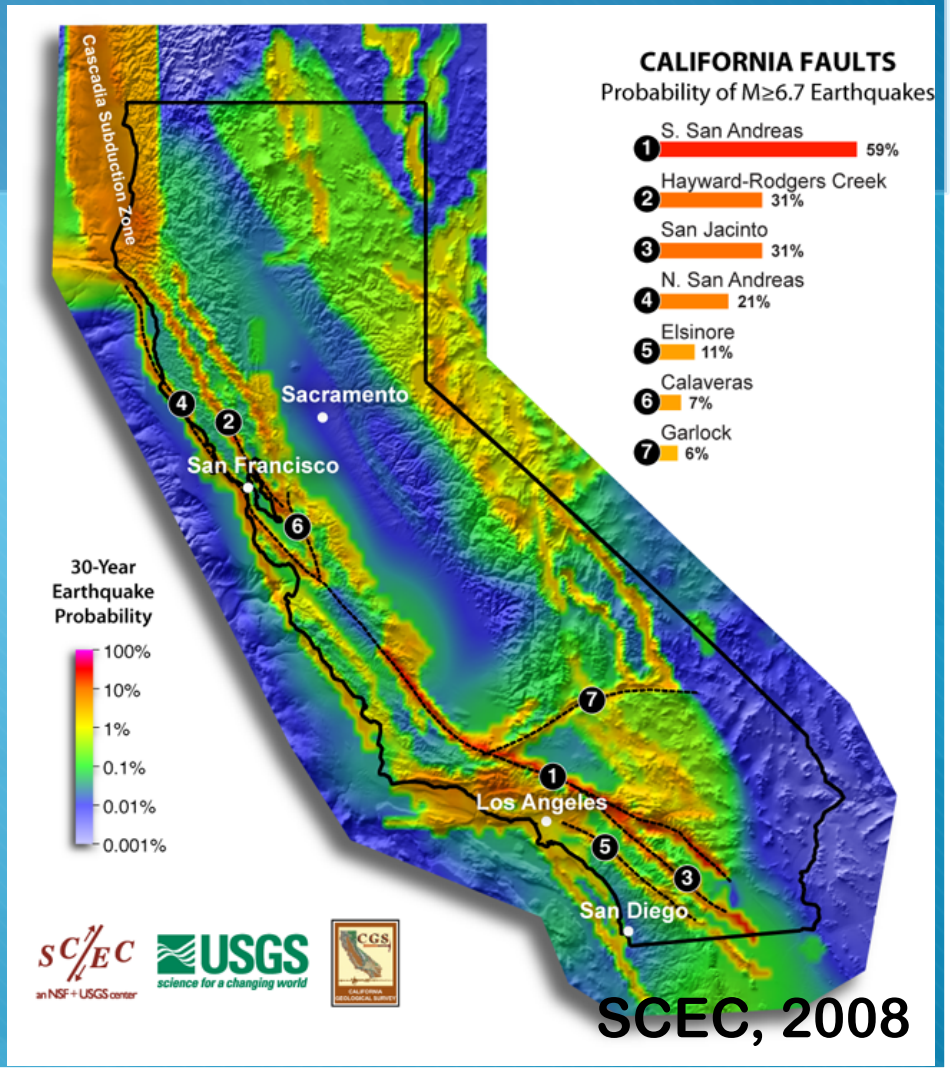
Seismology Basics: Waves

Seismic wave speeds:

P-wave: 5-7 km/s

S-wave: 3-4 km/s

S waves are slower – But cause the most damage

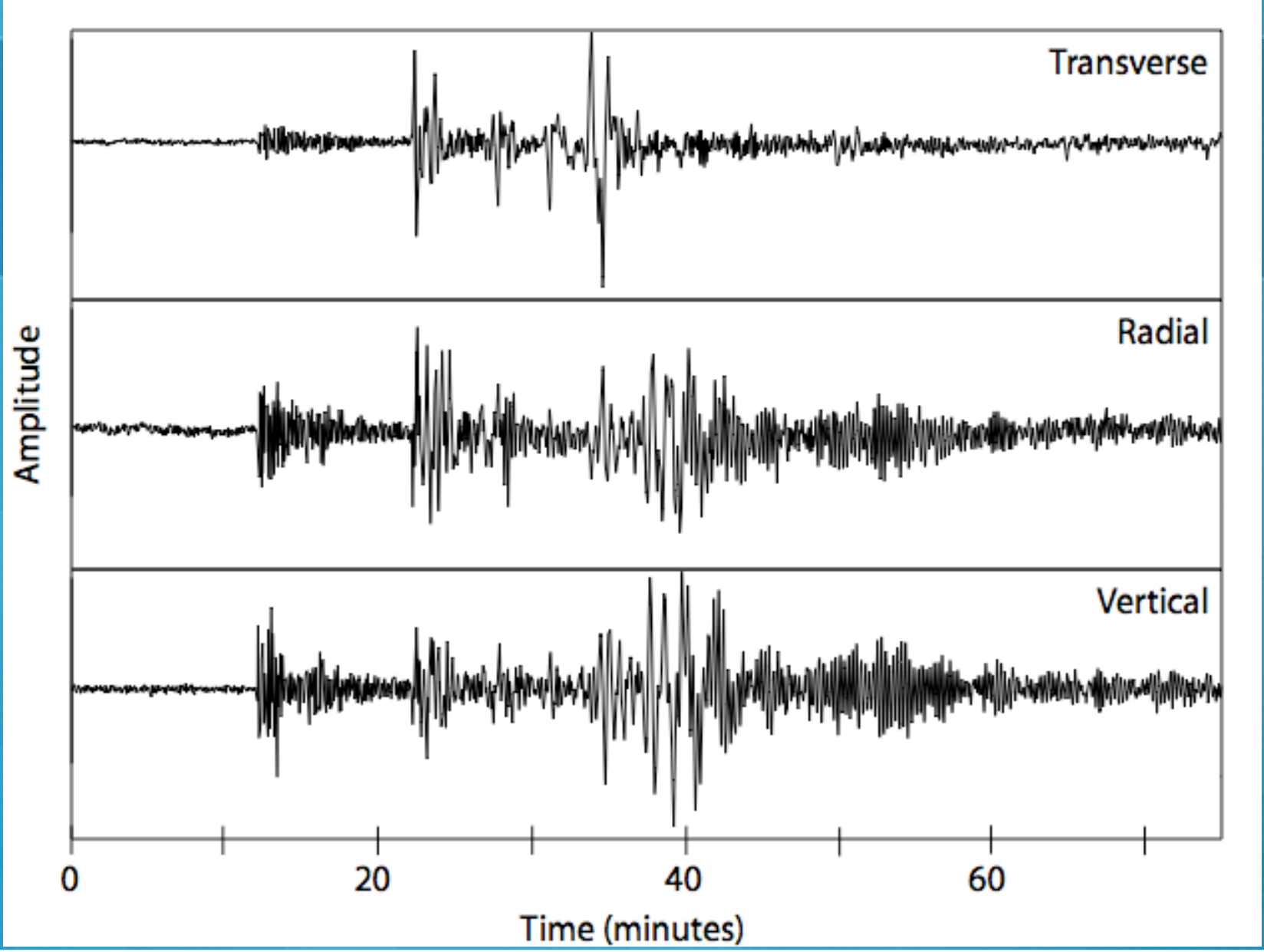


SCEC, 2008

Seismology Basics: Waves (cont'd)

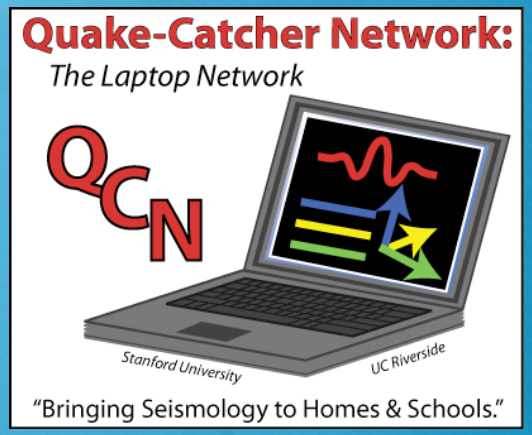
P-wave: First Arrival

S, Love, and Rayleigh Waves: More Damaging Arrivals



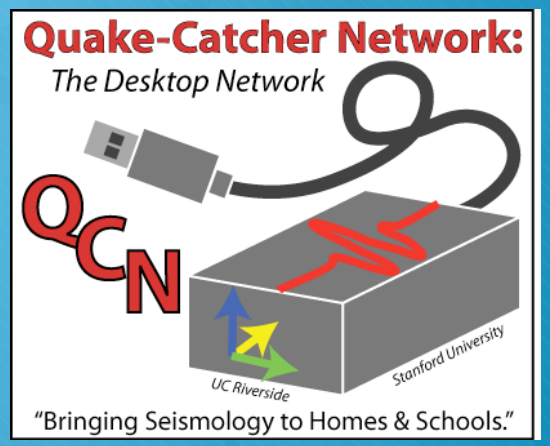
QCN: No-Cost Network

- Laptops With Sensors:
- HP, Apple, ThinkPad, Acer
- Very noisy data, not coupled to ground
- Sensors move location – track online
- Drawback: info on sensor detection and usage tough to get from manufacturers: only Apple Mac & ThinkPad supported by QCN



QCN: Low-Cost Network

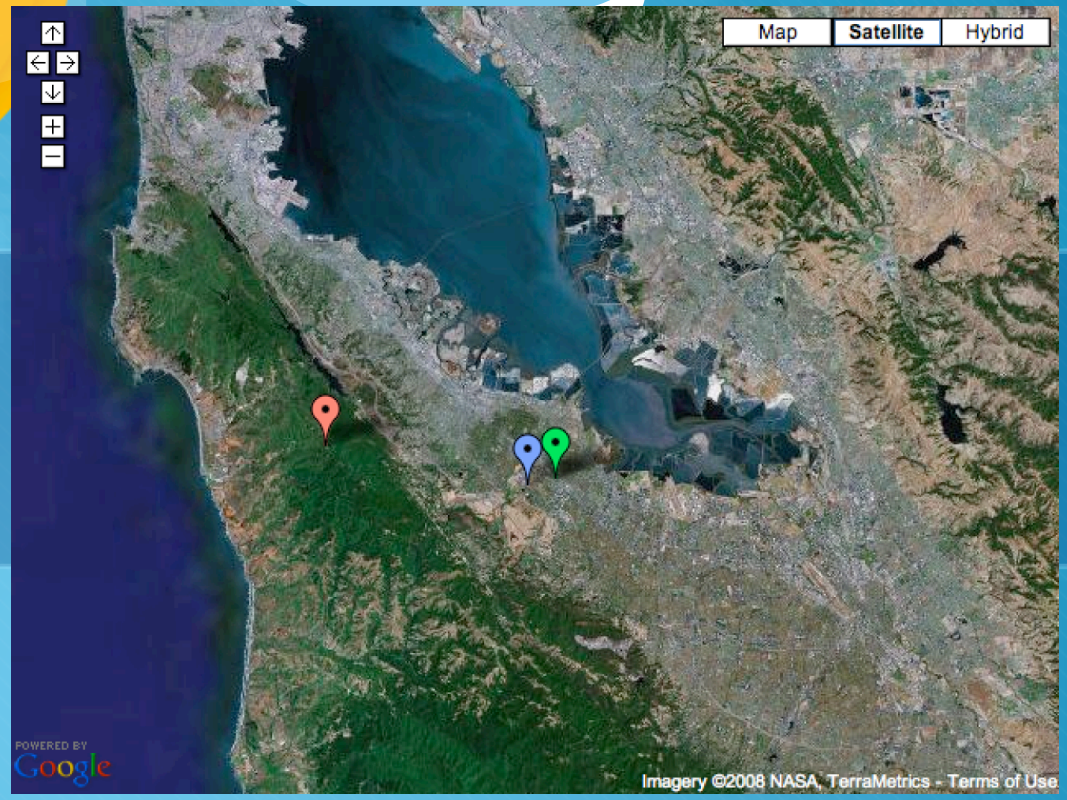
- Desktops with connected USB sensors
- Cost: $35-$150 per sensor
- In noisy environments (homes & businesses)
- Over time hardware is getting cheaper, sensitivity/features are increasing
- 5000 purchased for distribution (Paypal purchases etc):
- http://qcn.stanford.edu/learning/requests.php



QCN Challenges

- Better Sensors: More sensitivity = better science
- *Location:* always changing – track IP addresses on our website, user inputs their location lat/lng
- *Noise:* typing, bouncing on laps, slamming doors, running kids, …
- *Timing:* no GPS clock – synchronized to our servers via ntp (network time protocol)

Location



3-step Location System:

- Estimate location based on last known router (geoip)Often accurate within several kilometers
- Participants provide their "favorite five locations" using a Google Maps Interface.
- linked to IP or set a default location/address

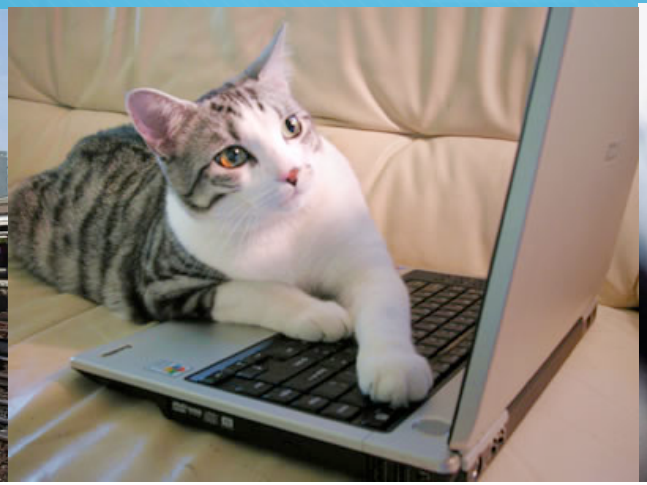
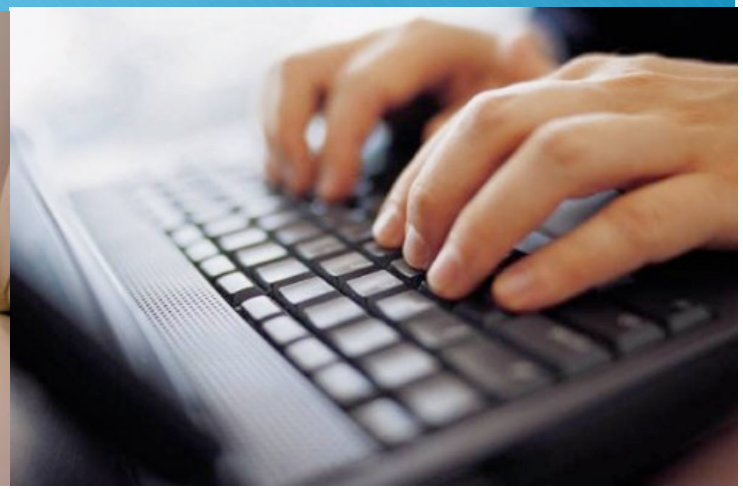
Future: a "Where Were You?" website

Select	Location Name (optional)	Latitude	Longitude	Net (IP) Addr	Set Net Addr	Clear Net Addr
	Home	34.0971731803043	-117.72793114185	76.170.119	Set Current	Clear
	Work	33.9745572764349	-117.32615232467	138.23.128	Set Current	Clear
					Set Current	Clear
					Set Current	Clear
					Set Current	Clear

Update info

Battling Noise

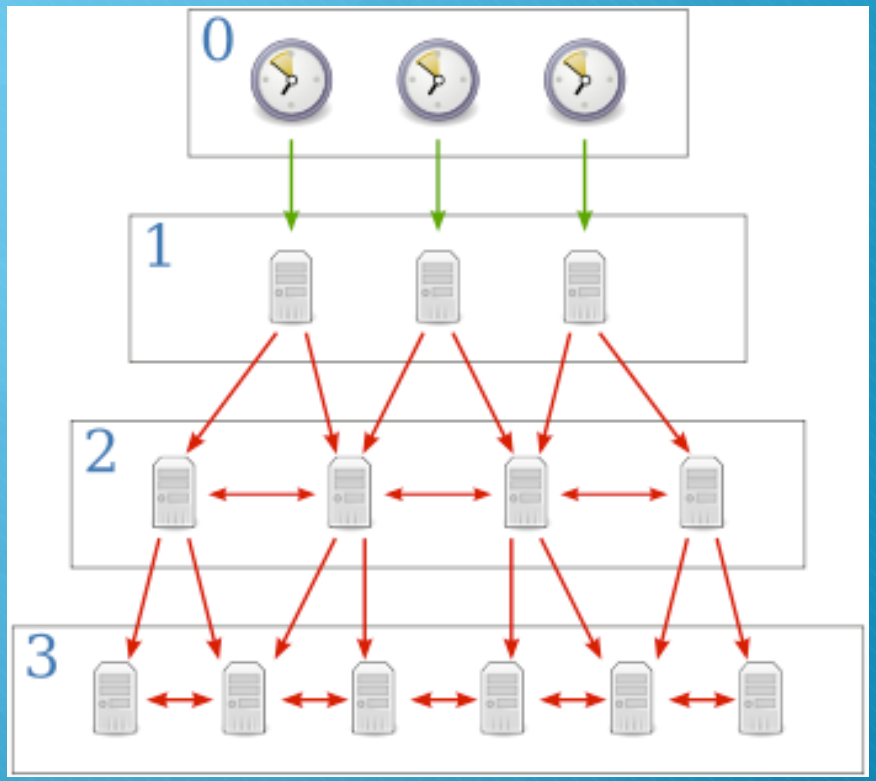
- Use a statistical triggering algorithm that self-adjusts for the noise level
- We only monitor sensors when the laptop is otherwise unused
- Use 1,000s-100,000s sensors, not 10s-100s
  - Rely on statistics of large numbers

Timing

- Network Time Protocol (NTP, 1985):
- Multi-tier system grounded to:
  - GPS Clocks
  - Atomic Clocks
  - Radio Clocks

Peer-to-peer method often provides better than 0.1 second accuracy, often +/- 20 msec.

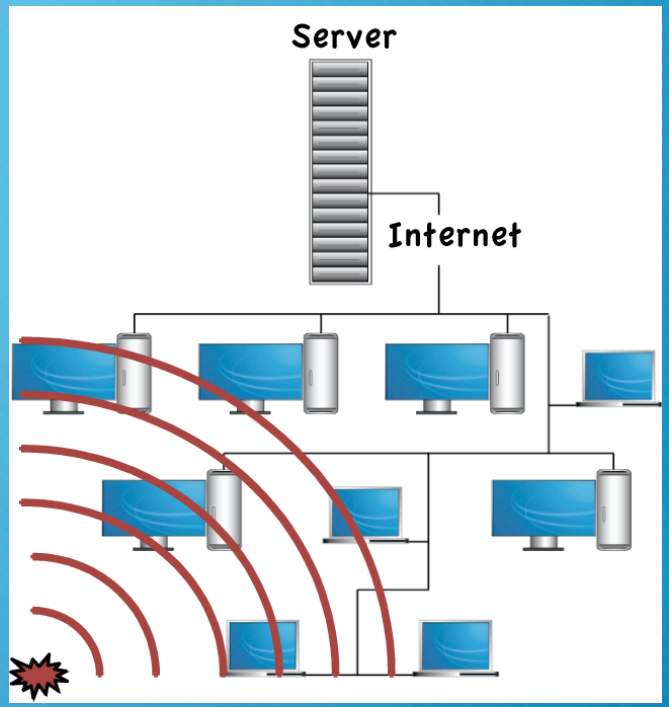


Frassetto et al. (SRL, 2003)

http://en.wikipedia.org/wiki/Network_Time_Protocol

Earthquake Detection

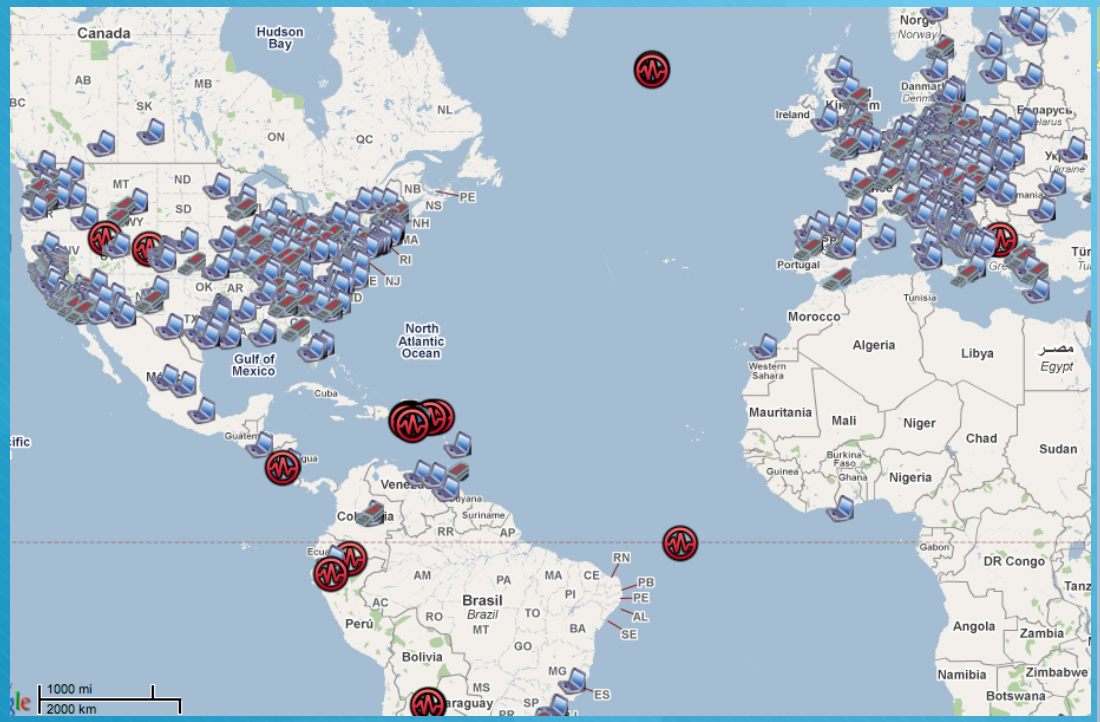
- Probable earthquake detection when the QCN receives many triggers from a region
- Otherwise just people bumping their laptops
- For big earthquakes:
  - only strong vibrations will be detected
  - Only large earthquakes will cause consistent triggers across a region of the network

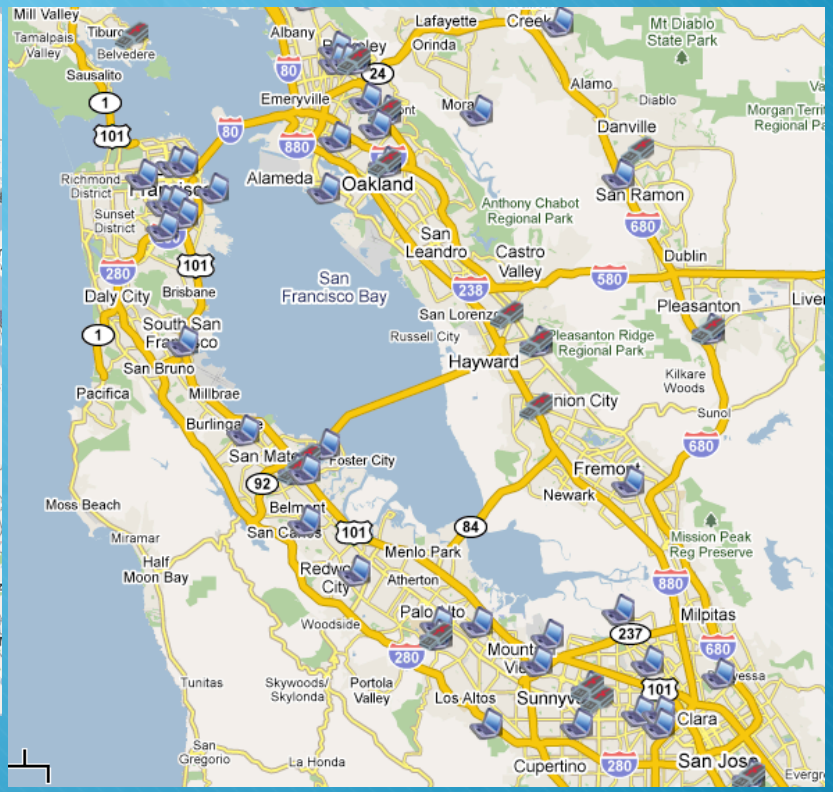


Rapid Response

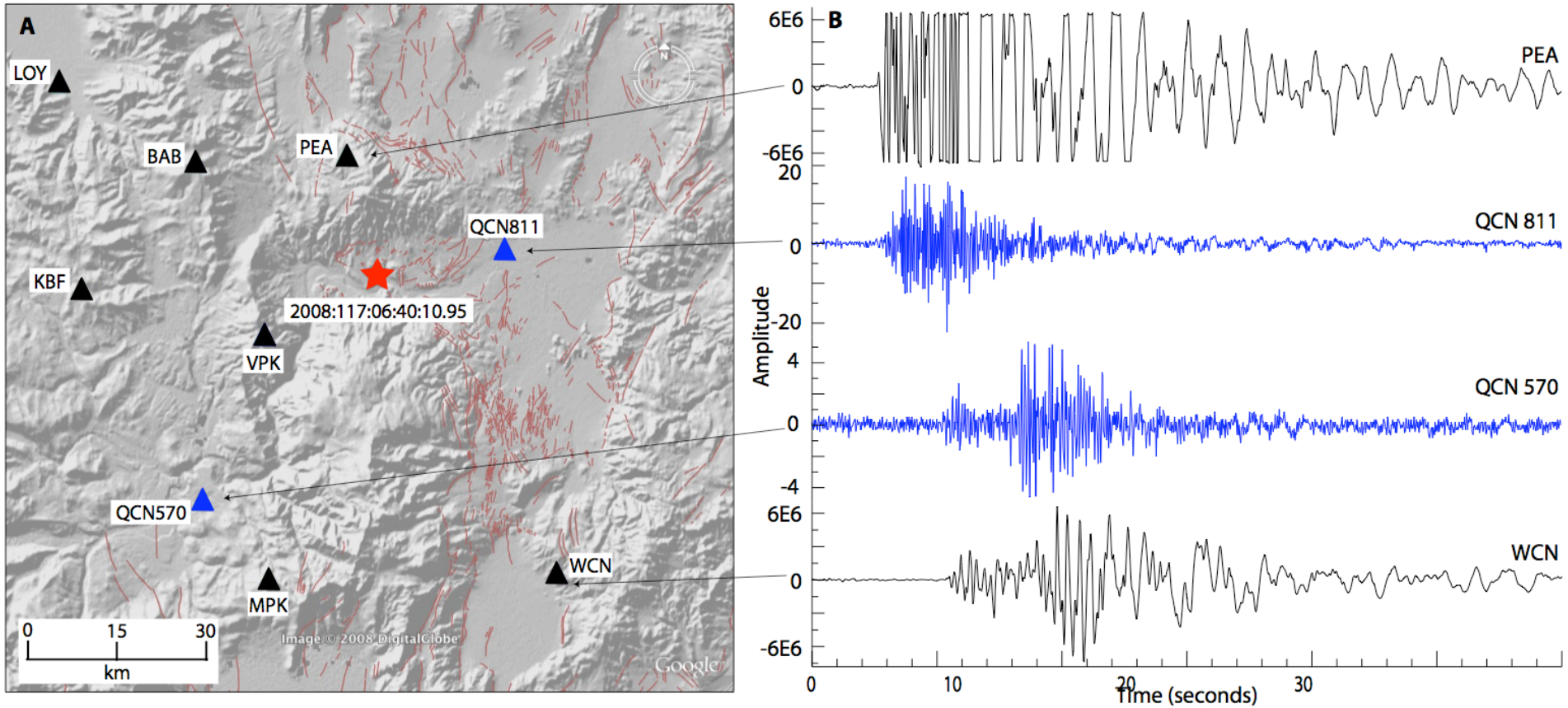
- Use a fast triggering algorithm
- Transfer statistics, not waveforms:
- Small XML "packet" that the BOINC client sends up instantly – seismic data (SAC files) can be requested for upload later via BOINC
  - Significance level, time, IP, and XYZ amplitudes.
  - With enough sensors, statistical models can work well.
  - Store waveform data temporarily on the sensor-computer.
  - Sensor-to-server in 100ms (LA region) to 5 s (around the world) depending on local network traffic/bandwidth.

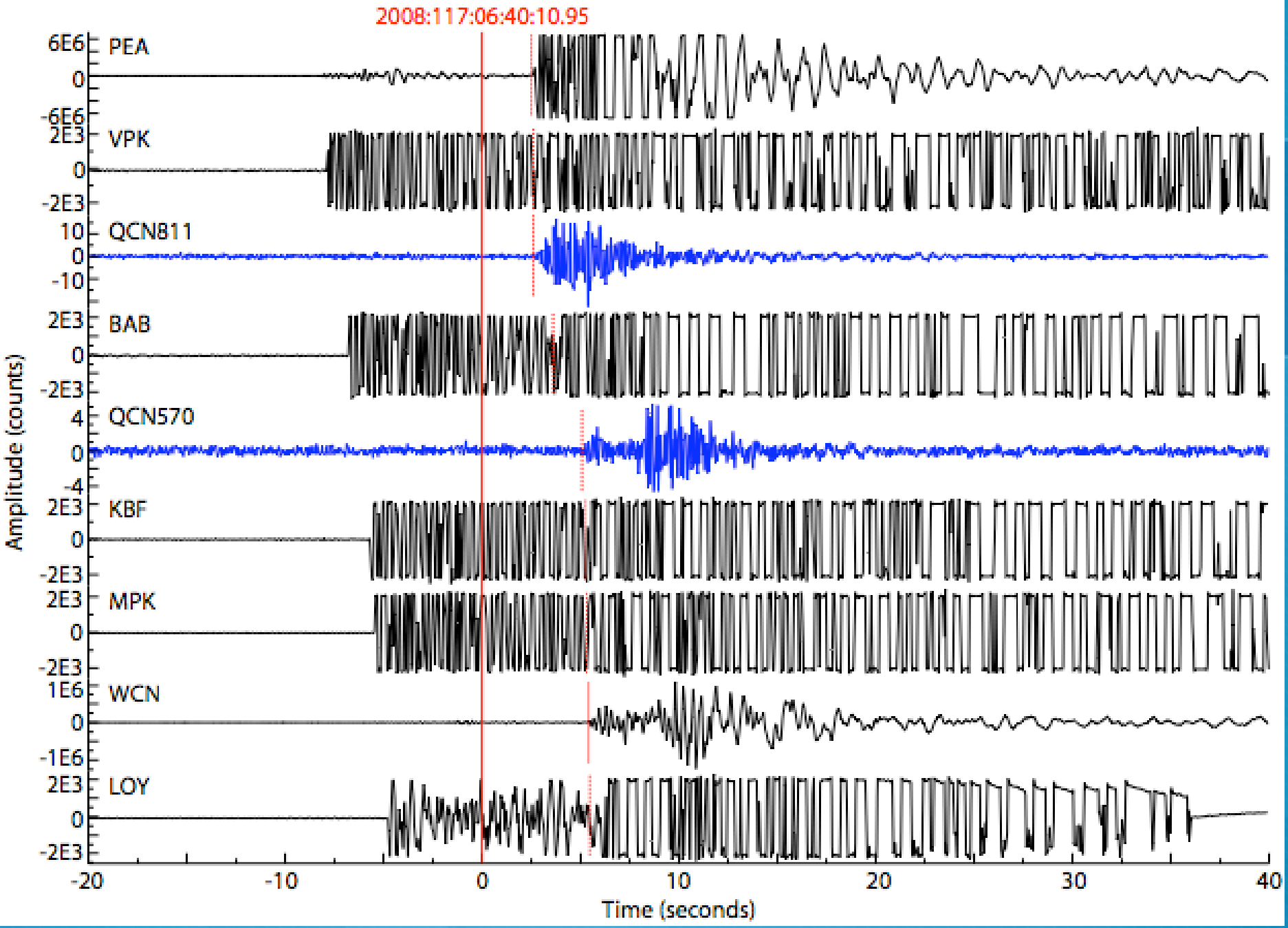
Network Status ~1000 (09/2009)





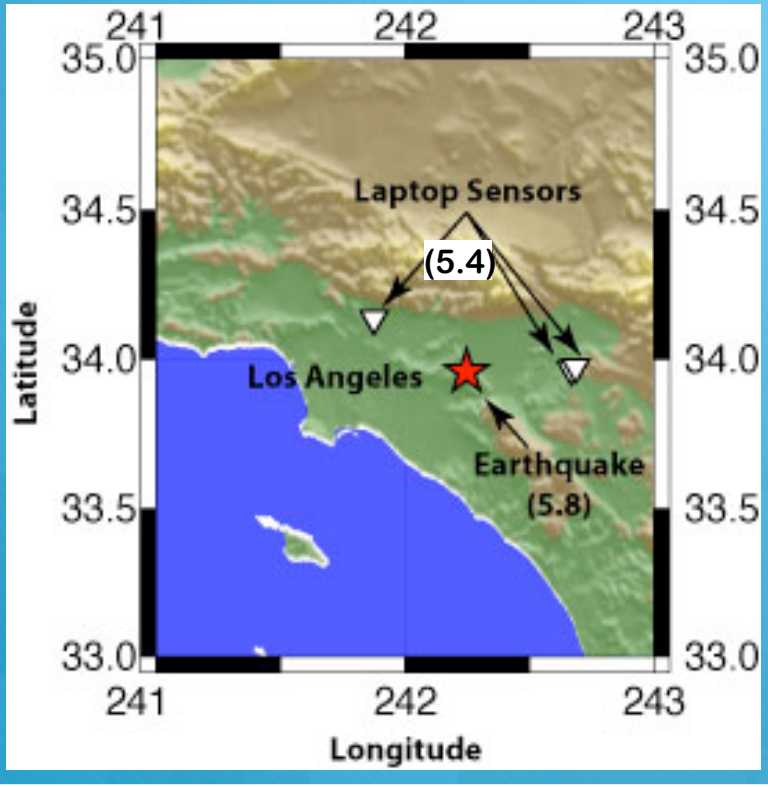
Reno Earthquakes Captured – 04/2008

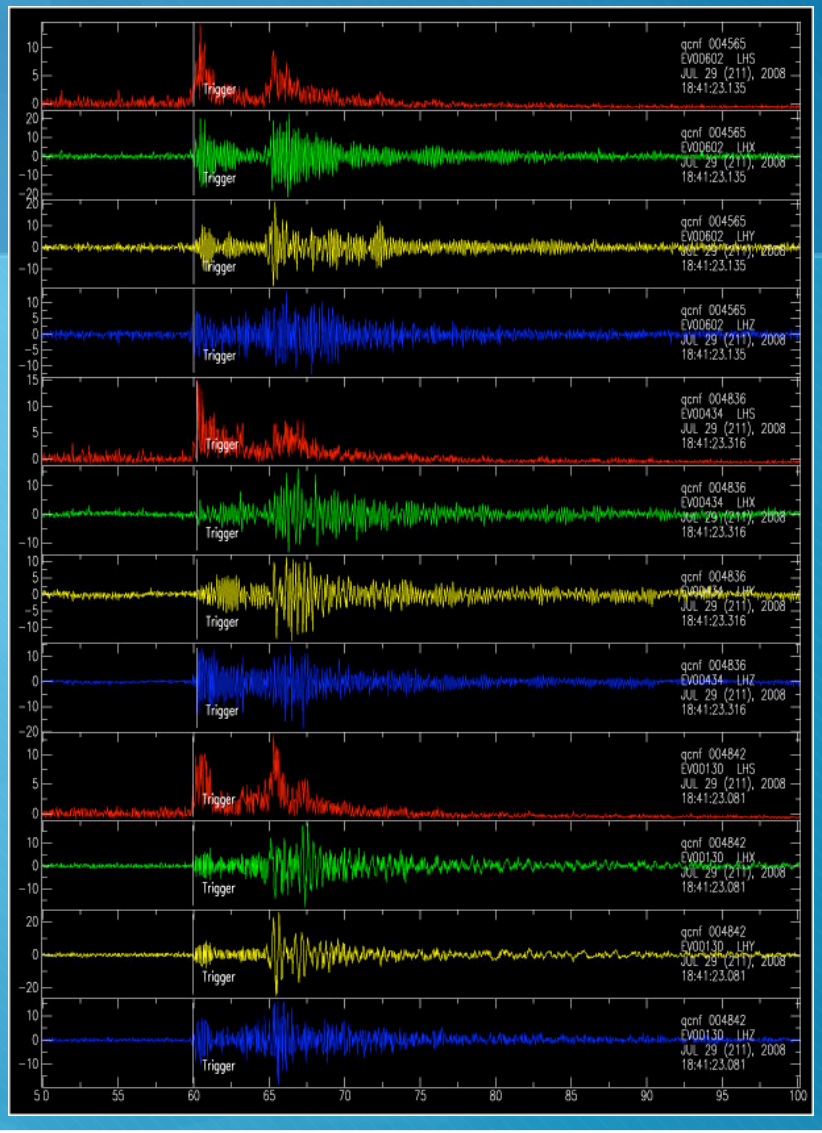



2008:117:06:40:10.95
PEA
VPK
QCN811
BAB
QCN570
KBF
MPK
WCN
LOY
6E6
0
-6E6
2E3
-2E3
10
-10
4
-4
1E6
-1E6
Amplitude (counts)
-20
-10
0
10
20
30
40
Time (seconds)

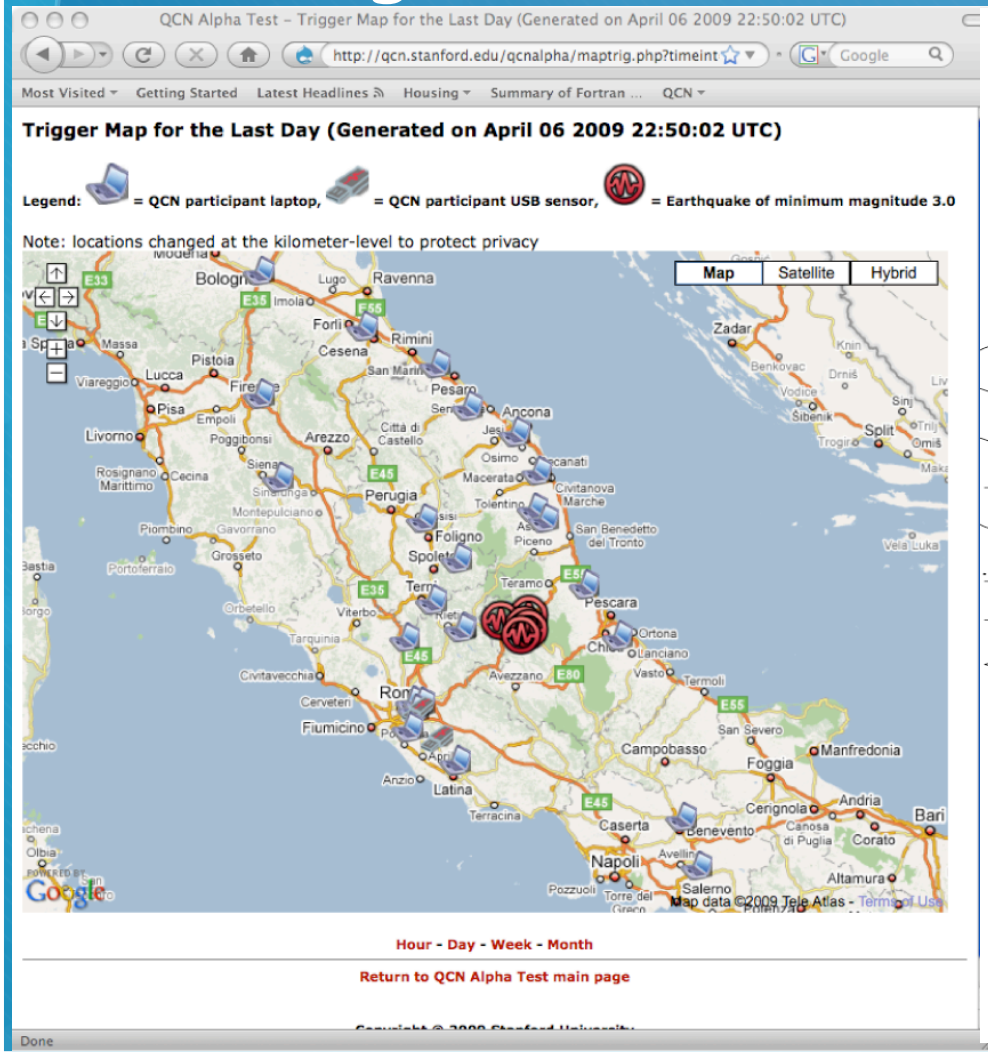
Detection, Location & Magnitude

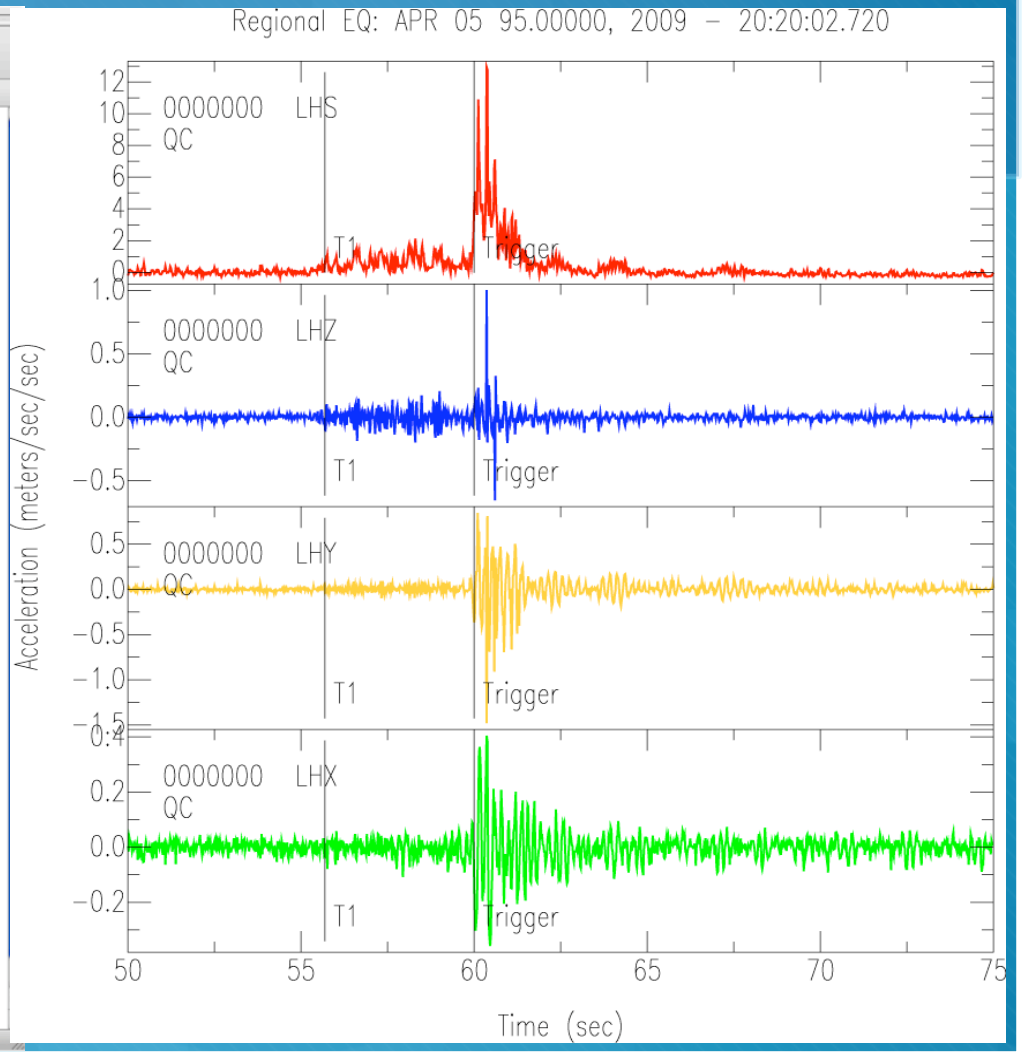
LA/Chino Hills Earthquake (07/2008)





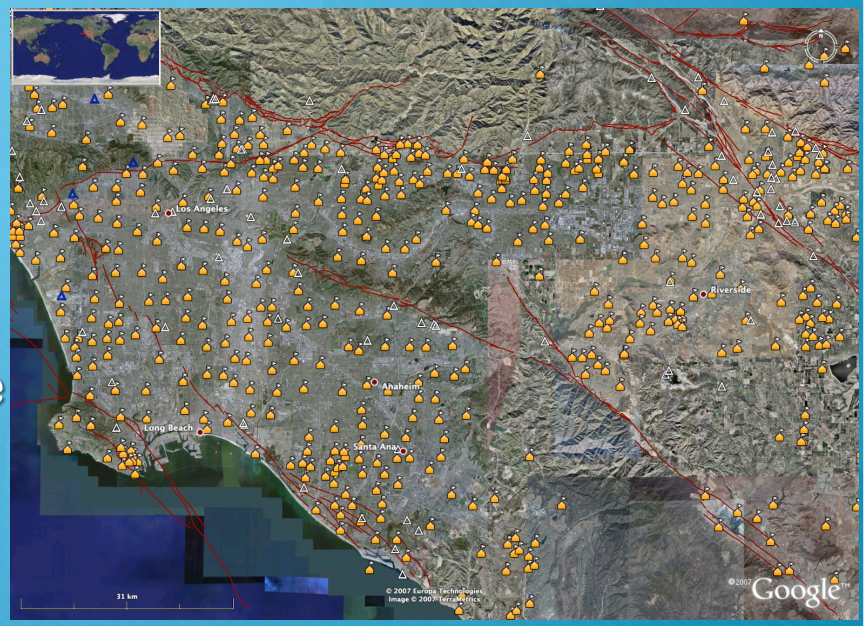
Italy 04/2009 – 3 sensors





Desktop Network

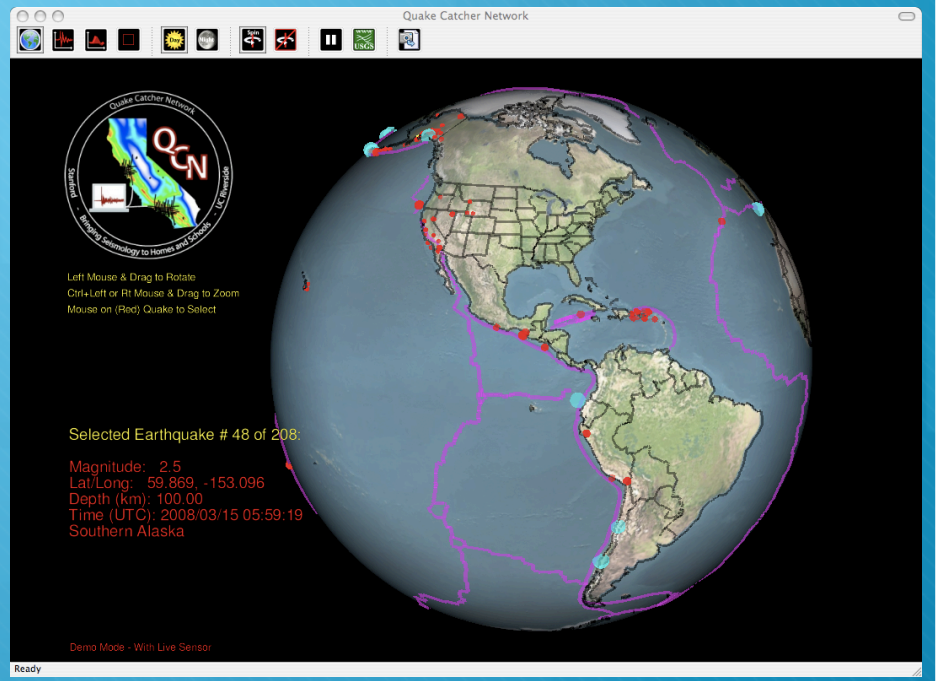
- With a USB sensor, any computer can be turned into a strong motion seismometer with QCN software
  - Schools can use the software to educate students about earthquake & seismology
  - School sensors can be distributed evenly with population



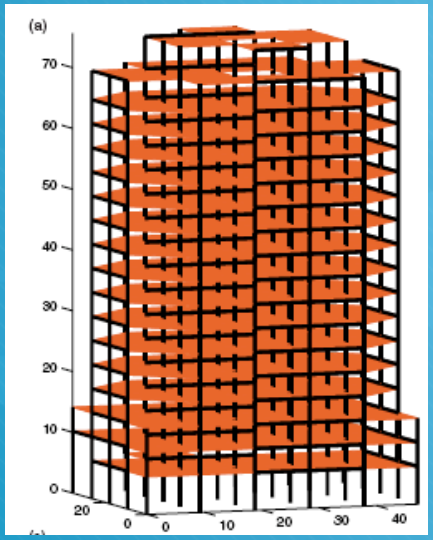
Subset of K-12 schools in LA basin

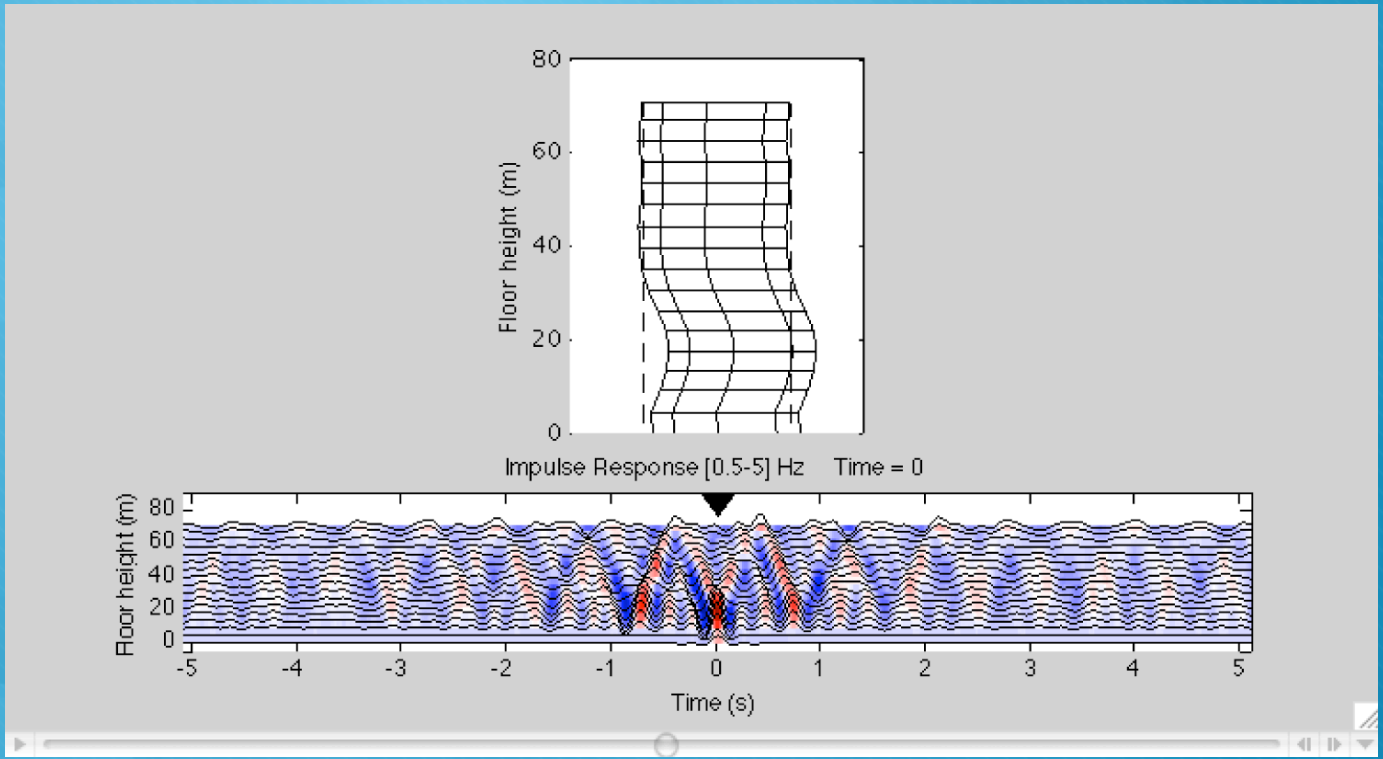
Educational Outreach

- What we provide:
  - Classroom Demo software.
  - Seismology related in-class activities.
  - Classroom USB Sensor.
  - Classroom BOINC Software.



Building Response: From Ambient Seismic Field (just need one week's worth of data)





Conclusion

- Volunteer computing is a proven resource for large earth science projects whether low or high-CPU requirements
- Exciting new avenues of approach – 64-bit, GPU, gaming consoles, multicore/multithreaded
- I would love to hear of any new ideas, suggestions, or proposals you may have, please email:

  carlc@atm.ox.ac.uk