Topix.net Weblog

News and information about Topix.net

| **Main** |

February 2005

Sun	Mon	Tue	Wed	Thu	Fri	Sat
		1	2	3	4	5
6	7	8	9	10	11	**12**
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

April 04, 2004

The Secret Source of Google's Power

Much is being written about Gmail, Google's new free webmail system. There's something deeper to learn about Google from this product than the initial reaction to the product features, however. Ignore for a moment the observations about Google leapfrogging their competitors with more user value and a new feature or two. Or Google diversifying away from search into other applications; they've been doing that for a while. Or the privacy red herring.

No, the story is about seemingly incremental features that are actually massively expensive for others to match, and the *platform* that Google is building which makes it cheaper and easier for them to develop and run web-scale applications than anyone else.

I've written before about Google's snippet service, which required that they store the entire web in RAM. All so they could generate a slightly better page excerpt than other search engines.

Google has taken the last 10 years of systems software research out of university labs, and built their own proprietary, production quality system. What is this platform that Google is building? It's a

Search

Search this site:

Archives

Recent Entries

distributed computing platform that can manage web-scale datasets on 100,000 node server clusters. It includes a petabyte, distributed, fault tolerant filesystem, distributed RPC code, probably network shared memory and process migration. And a datacenter management system which lets a handful of ops engineers effectively run 100,000 servers. Any of these projects could be the sole focus of a startup.

Speculation: Gmail's Architecture and Economics

Let's make some guesses about how one might build a Gmail.

Hotmail has 60 million users. Gmail's design should be comparable, and should scale to 100 million users. It will only have to support a couple of million in the first year though.

The most obvious challenge is the storage. You can't lose people's email, and you don't want to ever be down, so data has to be replicated. RAID is no good; when a disk fails, a human needs to replace the bad disk, or there is risk of data loss if more disks fail. One imagines the old ENIAC technician running up and down the isles of Google's data center with a shopping cart full of spare disk drives instead of vacuum tubes. RAID also requires more expensive hardware -- at least the hot swap drive trays. And RAID doesn't handle high availability at the server level anyway.

No. Google has 100,000 servers. [nytimes] If a server/disk dies, they leave it dead in the rack, to be reclaimed/replaced later. Hardware failures

Links

Topix.net
About Us
Team
Press
FAQ's
Feedback

Topix.net Favorites

Search Engine news
News on Google, Inc.
Blog news
Spam news
Palo Alto, CA local news
Weird

Interesting

Search Engine Watch
Traffick
ResourceShelf
Susan Mernit
Jeff Jarvis
John Battelle
Feedster
editorsweblog.org
WebTalk Radio Show
Syndic8
Micro Persuasion

Syndicate this site (XML)

Powered by
Movable Type

need to be instantly routed around by software.

Google has built their own distributed, fault-tolerant, petabyte filesystem, the [Google Filesystem](). This is ideal for the job. Say GFS replicates user email in three places; if a disk or a server dies, GFS can automatically make a new copy from one of the remaining two. Compress the email for a 3:1 storage win, then store user's email in three locations, and their raw storage need is approximately equivalent to the user's mail size.

The Gmail servers wouldn't be top-heavy with lots of disk. They need the CPU for indexing and page view serving anyway. No fancy RAID card or hot-swap trays, just 1-2 disks per 1U server.

It's straightforward to spreadsheet out the economics of the service, taking into account average storage per user, cost of the servers, and monetization per user per year. Google apparently puts the operational cost of storage at $2 per gigabyte. My napkin math comes up with numbers in the same ballpark. I would assume the yearly monetized value of a webmail user to be in the $1-10 range.

Cheap Hardware

Here's an anecdote to illustrate how far Google's cultural approach to hardware cost is different from the norm, and what it means as a component of their competitive advantage.

In a previous job I specified 40 moderately-priced servers to run a new internet search site we were developing. The ops team overrode me; they wanted 6 more expensive servers, since they said it would be easier to manage 6 machines than 40.

What this does is raise the cost of a CPU second. We had engineers that could imagine algorithms that would give marginally better search results, but if the algorithm was 10 times slower than the current code, ops would have to add 10X the number of machines to the datacenter. If you've already got $20 million invested in a modest collection of Suns, going 10X to run some fancier code is not an option.

Google has 100,000 servers.

Any sane ops person would rather go with a fancy $5000 server than a bare $500 motherboard plus disks sitting exposed on a tray. But that's a 10X difference to the cost of a CPU cycle. And this frees up the algorithm designers to invent better stuff.

Without cheap CPU cycles, the coders won't even consider algorithms that the Google guys are deploying. They're just too expensive to run.

Google doesn't deploy bare motherboards on exposed trays anymore; they're on at least the fourth iteration of their cheap hardware platform. Google now has an institutional competence building and maintaining servers that cost a lot less than the servers everyone else is using. And they do it with fewer people.

Think of the little internal factory they must have to deploy servers, and the level of automation needed to run that many boxes. Either network boot or a production line to pre-install disk images. Servers that self-configure on boot to determine their network config and load the latest rev of the software they'll be running. Normal datacenter ops practices don't scale to what Google has.

What are all those OS Researchers doing at Google?

Rob Pike has gone to Google. Yes, that Rob Pike -- the OS researcher, the member of the original Unix team from Bell Labs. This guy isn't just some labs hood ornament; he writes code, lots of it. Big chunks of whole new operating systems like Plan 9.

Look at the depth of the research background of the Google employees in OS, networking, and distributed systems. Compiler Optimization. Thread migration. Distributed shared memory.

I'm a sucker for cool OS research. Browsing papers from Google employees about distributed systems, thread migration, network shared memory, GFS, makes me feel like a kid in Tomorrowland wondering when we're going to Mars. Wouldn't it be great, as an engineer, to have production versions of all this great research.

Google engineers do!

Competitive Advantage

Google is a company that has built a single very large, custom computer. It's running their own cluster operating system. They make their big computer even bigger and faster each month, while lowering the cost of CPU cycles. It's looking more like a general purpose platform than a cluster optimized for a single application.

While competitors are targeting the individual applications Google has deployed, Google is building a massive, general purpose computing platform for web-scale programming.

This computer is running the world's top search engine, a social networking service, a shopping price comparison engine, a new email service, and a local search/yellow pages engine. What will they do next with the world's biggest computer and most advanced operating system?

Posted by skrenta at April 4, 2004 02:11 PM

Comments

Wow! That's the most fascinating blog entry I've read in a long time. Just one point of contention: I would be very surprised if Orkut is running on Google's massive distributed platform. Orkut appears to be implemented in ASP.NET, which would suggest it's running on Microsoft servers. I very much doubt Windows is running on Google's 100,000 core servers so I would assume that Orkut has it's own little cluster of Windows servers somewhere.

Posted by: **Simon Willison** at April 5, 2004 12:37 AM

I understand that google does have people with shopping carts full of disks going around replacing them and that they do not use white boxes but use 1U servers.

Does anyone have more definitine info?

Posted by: **Rob** at April 5, 2004 01:24 AM

Where are you getting this number for 100,000 servers? Everything I have read

has indicated 10,000

Posted by: **Brian** at April 5, 2004 01:30 AM

The 100,000 figure is from:

NY Times: The Coming Search Wars.

Previously I'd heard 10k figures, then later 50k figures. The latest is this 100k figure. Most recently I saw a craigslist job post hiring ops guys to run 10k servers, but I assume that is a datacenter's worth, and they have multiple datacenters.

Posted by: **Rich Skrenta** at April 5, 2004 01:34 AM

The story about leaving drives sitting bare on top of motherboards was from a talk Sergey gave at VA Research a few years ago. They had a single power supply which would run 2 little motherboards, lay the disks on top, and lay the whole mess on a tray.

Because everything was so fragile it was too difficult to get to dead servers, so they just left them in place in the rack. Sergey's talk had some great photos of these early racks but sadly I couldn't find them for my post.

After that I believe Google used **Rackable**, which can get 80 servers into the space of a single rack (normally 40 would be max). I don't know what they're using now.

Posted by: **Rich Skrenta** at April 5, 2004 01:38 AM

Last I heard, Google has about 40K servers. They used to have
the exposed motherboards on a shelf systems, with
switches nylon-strapped to the tops of the cabinets, and spaghetti wire everywhere. They are
now using I believe Rackable Systems servers
(www.rackable.com) in some of their datacenters.
On the old hardware, I heard the failure rate was
about 25%. I can't remember if those were 2 or 4
per shelf. The new hardware is (if my memory
serves) 1U but half depth servers, lined up in the
front AND back of the cabinets, with fans on top
sucking the hot air up and out. Google has a lot
staff just for datacenter ops, and more who do
sysadmin work. Those old shelf based systems threw
out a LOT of heat.

Posted by: **dude** at April 5, 2004 01:48 AM

Google might win millions of users for Gmail
in the first year and outwit the theories of
market and placement even, but so much
diversification and Spam might give it a
tough time than Yahoo and Microsoft
together. And Google founders surely don't

want an IPO at the cost of their freedom of decision-making and innovation. What I am skeptical about Google's scale of growth is the question whether there's ever been such a massive organization who was backed by Academic Research like Google?

Posted by: **Ejaz Asi** at April 5, 2004 02:34 AM

I'd check out **this article** as well, where the pure mail dynamics of the thing are discussed at length. It looks feasible even on conventional platforms.

Posted by: **Michael** at April 5, 2004 03:03 AM

Check out this IEEE paper for older information: http://discuss.fogcreek.com/redirect.asp?http://www.computer.org/micro/mi2003/m2022.pdf

Posted by: **Raju Varghese** at April 5, 2004 03:29 AM

Is this different from what **Opsware** is doing?

Posted by: **Sean Brunnock** at April 5, 2004 03:37 AM

Fascinating read! Of course, the paranoid part of me wonders if Google is going to become Skynet with all that power :P

Posted by: **Ryan VanderMeulen** at April 5, 2004

06:51 AM

hello,

what is skynet ?

Posted by: **octopuce** at April 5, 2004 08:58 AM

Octopuce -- Skynet is a fictional rogue AI, part of the backstory in the US sci-fi movie "The Terminator".

Paranoia or no, the Google platform does look like promising grounds for future artificial intelligence work. How many servers would you need to emulate the human brain?

Posted by: **matt** at April 5, 2004 09:13 AM

The 100,000 number is accurate.

Posted by: **Jeremy Zawodny** at April 5, 2004 09:29 AM

Amazing figures and enjoyable post. Other thing is that the Google mindset is a clear indication of the way its engineers are approaching their solutions: Instead of wondering whether hardware is cheap or expensive, the Google team focused in the cost of the process. It implies a jump from the quantification of assets to the qualification of those.
While obvious, what does it mean in an

economy still dominated by the amount of goods produced and sold?

Posted by: **Camilo** at April 5, 2004 10:20 AM

Interesting post, I had been thinking about similar ideas when I posted over the weekend.
(http://www.xxeo.com/)

Also, the story about the 2 motherboards is inaccurate. It was 4 motherboards per tray with
a plexiglass sheet on top with drives on top of that.

Dru

Posted by: **Dru Nelson** at April 5, 2004 11:12 AM

Thanks for one of the most interesting blog articles I've seen in a while! A very good read...

InfoWorld: "For instance, Rackable customers include Yahoo and Google, with their Web farms and search engines" indicates that Google uses Rackable. Although it doesn't say anything about how much/which Google systems use Rackable. Searching Google for **rackable+google** gives many hits...

Posted by: **John Magnus** at April 5, 2004 11:17 AM

Eyes and mouth wide open :o

Posted by: **Teucer** at April 5, 2004 11:58 AM

"This computer is running the world's top search engine, a social networking service, a shopping price comparison engine, a new email service, and a local search/yellow pages engine. What will they do next with the world's biggest computer and most advanced operating system? "

… They also running Google News, which seems to be totally independent of their search engine - since the News site indexes pages that never show up in the regular Google search (probably due to robots.txt limitations of news sites ?)

~sumedh

Posted by: **Sumedh Mungee** at April 5, 2004 12:09 PM

So how many entries in the Supercomputer Top 500 would one have to combine to compare it to the approximate power of Googles combined CPUs?

Posted by: **Isotopp** at April 5, 2004 01:16 PM

Very impressive writeup…should be a HBS case…

Posted by: **Michael Porter** at April 5, 2004 05:07 PM

Re: Simon Willison's comment about being run on Windows.

It's not:
http://uptime.netcraft.com/up/ graph/?host=orkut.com

Posted by: **Erik** at April 5, 2004 06:28 PM

Netcraft is quite amusing.

Posted by: **Ryan** at April 5, 2004 08:38 PM

Great article. As far as those rackable servers, those might be used in some of their adwords and adsense servers. I would imagine that those servers need to be more robust than the the solid search/spidering servers since they deal directly with paying clients. Just a guess though.

Posted by: **werty** at April 5, 2004 09:32 PM

just answering an above question:

"How many servers would you need to emulate the human brain?"

you would need at least 100 billion (100,000,000,000) computers to even get close to a human brain. google isn't quite there yet..

Posted by: **ak** at April 5, 2004 10:03 PM

Fascinating stuff.

>>I would assume the yearly monetized value of a webmail user to be in the $1-10 range.

It's way higher that that. Think AdWords.

Posted by: **Peter** at April 5, 2004 10:42 PM

Google's control over information is going further and further. We use his services every day. He knows us, our questions, our desires, our needs, how we live. Everything is stored into its database. Information is power, commercial power, political power. Who's controlling Google ?

http://www.google-watch.org/

Posted by: **Jeffrey Lebowsky** at April 6, 2004 12:44 AM

Google is getting so powerful that if it does not turn to evil, it could find itself nationalized because unregulated, it is a threat to national security. **This slashdot post** outlines one of its potential government uses.

Google is in danger, and not from M$. Something this large and powerful that has an impact on the lives of billions of people simply will not be allowed to be independently run for the good of mankind.

Posted by: **Owthright** at April 6, 2004 02:53 AM

"Something this large and powerful that has an impact on the lives of billions of people simply will not be allowed to be independently run for the good of mankind."

Are you telling me that you trust politicians more than Google? Why?

Posted by: **aaron wall** at April 6, 2004 04:05 AM

Googlewatch is pretty kooky, as is the guy behind it.

http://www.google-watch-watch.org/

Posted by: **Shane** at April 6, 2004 05:10 AM

just answering an above question:

> "How many servers would you need to emulate the human brain?"

> you would need at least 100 billion (100,000,000,000) computers to even get close to a human brain. google isn't quite there yet..

I think you mixed up two things: the human brain consists of about 10^11 neurons with probably 10^13 neural connections. I think it should be possible to simulate more than one neuron with a single computer. Plus: the processing speed in the brain is much more lower, than in computers. But it is

vrey difficult to achieve a interconnectivity of 10^13 connections between the systems. However, you can surely say, google manages more information today than a single person can.

Posted by: **Mika** at April 6, 2004 07:11 AM

The argument that "Google is getting so powerful that if it does not turn to evil, it could find itself nationalized because unregulated, it is a threat to national security. This slashdot post outlines one of its potential government uses.

Google is in danger, and not from M$. Something this large and powerful that has an impact on the lives of billions of people simply will not be allowed to be independently run for the good of mankind."

actually applies 1000 times greater to Microsoft than Google. Everyone can stop using Google tomorrow and use AltaVista or Teoma and not suffer much. Everyone CANNOT stop running Windows tomorrow.

Posted by: **Jorge** at April 6, 2004 07:36 AM

Am I the only one who thought of Borges' Library of Babel when reading this post?

Posted by: **Terry** at April 6, 2004 07:52 AM

I think we can safely ignore (snicker) google-watch.

Hard to hear what the guy has to say over the rustling of his tinfoil hat anyway

Posted by: **John Kenneth Fisher** at April 6, 2004 08:08 AM

While obvious, what does it mean in an economy still dominated by the amount of goods produced and sold?

Google's "product" is their index, and their ability to use that index across a variety of services (search, comparison shopping, e-mail). I'm guessing that, now that they've perfected the ability to swap out and integrate low-cost CPUs and harddrives seamlessly, discrete servers are considered to be consumables or raw materials for all intents and purposes. Get 'em cheap, use 'em up, swap 'em out.

Posted by: **Ryland** at April 6, 2004 08:44 AM

Perhaps Google is the predecessor to the omnipotent core AI presented in Dan Simmon's Hyperion/Endymion novels?

Posted by: **Conal** at April 6, 2004 09:16 AM

One of the best blogs i have read about google and Gmail .The GFS was really intereting .

Is this equivalent to skynet , intresting thought ?

Posted by: **Abhilash M S** at April 6, 2004 11:53 AM

Hey. Great article… I really enjoyed reading it.

At rackable.com they have a box "who uses rackable?" and if you wait long enough you see the google logo.
Just thought I would mention that incase somebody else didn't clarify it since somebody said they didn't know what they used now.

-gonffen

Posted by: **gonffen** at April 6, 2004 12:45 PM

100,000 servers. Wow! Does anyone here know of any pics of the datacenter(s) that these things are stored in? And how the heck do they admin that many servers. I'm sure glad I'm not responsible for keeping 100,000 servers up and running. It gives me a headache just thinking about it.

Posted by: **Scott Johnson** at April 6, 2004 02:10 PM

> "How many servers would you need to emulate the human brain?"

> you would need at least 100 billion (100,000,000,000) computers to even get close to a human brain. google isn't quite there yet..

> I think you mixed up two things: the human brain consists of about 10^11 neurons with probably 10^13 neural connections.[snip]

Its also far from clear that just simulating neurons and their connections can approximate the human brain. For instance, I just read the cover article in the latest Scientific American talking about new research into glial cells which seems to indicate they might be more important to the brain's function that originally thought. I forget the numbers, and don't have the magazine in front of me, but I think it said there was something like 10 glial cells/ neuron. So you might need add another zero at least, along with more complicated connections.

Posted by: **Adam** at April 6, 2004 03:54 PM

Interesting posts everyone. There seems to be an AI thread, a server/datacenter thread, a google-future thread, and a google software thread all running together here.

Regardless, I would like to point out something about GMail in relation to spam.

"This computer is running the world's top search engine, a social networking service, a shopping price comparison engine, a new email service, and a local search/yellow pages engine."

Now, with such direct-access to so much information about *how* pieces of the web

relate to one another, wouldn't google be able to very effectively filter spam right from the beginning? Plus, with a "delete spam" option, google could harness individuals as "nodes" that indentify spam.

In conclusion, I think google already has a clear picture of spam, one which would rapidly be brought into sharper focus after a short period of GMail operating. Of any organization, they have the best information and technology with which to eliminate spam; and this is a service they could turn around and sell again.

Posted by: **Alexander Micek** at April 6, 2004 04:36 PM

Heck, I agree with Simon. This really is one of the most fascinating Blog entries that I've ever read. Even better than Simon's Blog at SitePoint. ;)

This sure sheds some light into what's cooking at Google & its already becoming a force to reckon with. Wotcher there Microsoft!! :D

Posted by: **Amit Gupta** at April 6, 2004 05:11 PM

Great, Great article.

Thanks

Posted by: **Tom** at April 6, 2004 05:30 PM

As Alexander stated, spam would be a non issue for Google. It's be nice if they'd release some of that information to the people who are filtering now (if they haven't already).

As for AI: if Google ever learns to build itself physically, I'm building a bunker and going into hiding, because at that point… Well, we would have skynet. It's scary thought, and to tell you the truth it gets into another article I read recently, somewhere, describing the ultimate virus.

The ultimate virus is one that can lay undetected on it's own for long long periods of time, learning the patterns of the virus scans, of other virus's, of routes across the web. It learns our behaviours, our methods, our actions. When it knows enough, it sends out infinitely weaker versions of itself such that it can probe for more knowledge on how it will be resisted (or not in the case of success), but in a way that the core of the virus won't be detected. Then, when the time is right, it unleashes the real version of itself while keeping a slightly modified version of itself in a few protected (not human touched) systems, this way, if for some reason, the virus get's stopped, it can start the process over yet again, and wait just long enough to try again.

Now I don't know about you, but to me, the first few steps of that sounds like Google. Sending out smaller versions of itself (google toolbar, various websites), learning (tracking us), and hiding itself well (who'd have thought of google as a virus?)

That said, it's scary, but a very good investment.

As for the service of Gmail… I will always run my own mail server.

Posted by: **JJ Doughboy** at April 6, 2004 07:47 PM

excellent eye opening reading ! many thanx everyone !

Posted by: **David Bubenicek** at April 6, 2004 07:58 PM

As the planet-wide GOOGLE computer continues to grow in size and sophistication, it finally receives a new name more descriptive of its impressive capabilities:

"O DEEP THOUGHT Computer," he said, "the task we have designed you to perform is this. We want you to tell us …" he paused, "… the Answer!"
"The answer?" said DEEP THOUGHT. "The answer to what?"
"Life!" urged Fook.
"The Universe!" said Lunkwill.
"Everything!" they said in chorus.
DEEP THOUGHT paused for a moment's reflection.
"Tricky," he said finally.
"But can you do it?"
Again, a significant pause.
"Yes," said DEEP THOUGHT, "I can do it."
"There is an answer?" said Fook with breathless excitement.
"A simple answer?" added Lunkwill.
"Yes," said DEEP THOUGHT. "Life, the

Universe, and Everything.
There is an answer. But," he added, "I'll have to think about it."

(from The Hitchhiker's Guide to the Galaxy, by Douglas Adams)

Posted by: **Ralph Dratman** at April 6, 2004 11:01 PM

Also see this very interesting lecture of Urs Hoelzle at Washington University (video stream at http://www.uwtv.org/programs/displayevent.asp?rid=1680) in November 2002.

It really is worth viewing!

"The Google Linux Cluster

Google's Linux cluster currently processes over 150 million queries a day, searching a multi-terabyte web index for every query with an average response time of less than a quarter of a second, with near-100% uptime. In this discussion, Google Fellow Urs Hölzle will describe the software and hardware infrastructure that makes this performance possible, as well as provide an overview of the main problems facing a web search, software architecture, servers and compact rack hardware designs."

Posted by: **mrgee** at April 6, 2004 11:04 PM

Does this have anything to do with google going public soon?

Posted by: **matt** at April 7, 2004 12:03 AM

Am I the only person who thinks this is basically a bunch of b.s. - give me 10 or 20 quality engineers and I can build a clustering system similar to Google's within a year.

A robust clustering system is not a real competitive advantage for Google versus Microsoft and Yahoo. I think AOL and Google should merge - AOL needs creativity and Google needs distribution, and fast.

Orkut's definitely on Win2k - its using pages with the extension .aspx and that doesn't run on Unix. Probably the first page is served off of Linux - not sure how else Netcraft is getting that result.

Posted by: **wag** at April 7, 2004 12:44 AM

.aspx *does* run on Linux (see http://www.go-mono.com/), but I think they're running it on Windows

Posted by: **Mauricio** at April 7, 2004 02:54 AM

Impressive information about what is becoming the King of the giants of computing…. Very good, thanks for this post…

Posted by: **Metropolis** at April 7, 2004 03:58 AM

Thanks - that is a great article!

Posted by: **Greasychipbutty** at April 7, 2004 05:08 AM

"I used to think that the human brain was the most awesome thing in the universe; but then I realized, look what's telling me that."

--Emo Philips

Posted by: **Guy** at April 7, 2004 08:44 AM

it seems like orkut.com is hosted on linux

netcraft.net says: "The site orkut.com is running orkut on Linux"

Posted by: **andi** at April 7, 2004 08:56 AM

How come when you do a search in Google for "search engine code" you get absolutely no search results?

How many search engines are there?

Posted by: **katmak** at April 7, 2004 10:34 AM

I'm new to all this stuff adnd was wondering is there nay free service to check out. I'm wanting to add some stuff to my web site to better communicate with everyone.

Posted by: **PsychicReader** at April 7, 2004 11:01

AM

Some credit for the performance of this architecture is owed to cheap gigabit ethernet on copper, i.e. low latency communication between nodes.

Harvesting the public network's knowledge and communication is slow by definition, but entity correlation for resale takes place on the high bandwidth, low latency, fixed cost, private network.

Posted by: **dotpeople** at April 7, 2004 11:03 AM

If Googlewatch is too kooky, here's a (hopefully) more balanced view (and a GoogleAnon, er, anonymizer) :-

http://www.imilly.com/google-cookie.htm

SkyNet and Deep Thought? Pah! How about :-

www.google.com/search?q="Is there a God?"

And for those who don't recognise the question :-

"Answer," from Angels and Spaceships, by Fredric Brown (1954).

"Dwar Ev ceremoniously soldered the final connection with gold. The eyes of a dozen television cameras watched him and the subether bore through the universe a dozen

pictures of what he was doing.

He straightened and nodded to Dwar Reyn, then moved to a position beside the switch that would complete the contact when he threw it. The switch that would connect, all at once, all of the monster computing machines of all the populated planets in the universe--ninety-six billion planets--into the supercircuit that would connect them all into the one supercalculator, one cybernetics machine that would combine all the knowledge of all the galaxies.

Dwar Reyn spoke briefly to the watching and listening trillions. Then, after a moment's silence, he said, "Now, Dwar Ev."

Dwar Ev threw the switch. There was a mighty hum, the surge of power from ninety-six billion planets. Lights flashed and quieted along the miles-long panel.

Dwar Ev stepped back and drew a deep breath. "The honor of asking the first question is yours, Dwar Reyn."

"Thank you," said Dwar Reyn. "It shall be a question that no single cybernetics machine has been able to answer."

He turned to face the machine. "Is there a God?"

The mighty voice answered without hesitation, without the clicking of single relay.

"Yes, now there is a God."

Sudden fear flashed on the face of Dwar Ev. He leaped to grab the switch.

A bolt of lightning from the cloudless sky struck him down and fused the switch shut."

;)

Posted by: **Milly** at April 7, 2004 01:08 PM

"search engine code" returned about 2,960 results or me actually…

Great article. Google is slowly becoming the monarch of virtual information.

The way it watches us, what we say, what we like, what we search for, what we buy, what we are interested in, it's not far from the truth to assume AI could be a project they're developing.

Look at their servers: they self configure themselves upon initial bootup to decrease the amount of manual labor. That's AI for you.

Posted by: **Brendyn** at April 7, 2004 02:27 PM

Google: The Microsoft of the 21st century.

I am afraid. I am very afraid.

Posted by: **Seth Finkelstein** at April 7, 2004 02:51 PM

FYI - we have a Google tracking page here at Topix.net:

www.topix.net/com/google

Posted by: **Rich Skrenta** at April 7, 2004 03:48 PM

i will believe it when i stop seeing this:

We apologize for the inconvenience, but we are unable to process your request at this time. Our engineers have been notified of this problem and will work to resolve it. Please note that using your browser's back button in AdWords can increase the likelihood of errors. If you think this was the cause of your error, please try again without using the back button.
----
i've been seeing it almost daily when i try to make changes to my adwords acct.

Posted by: **bobbyjimmy** at April 7, 2004 04:27 PM

Google is already as smart as Deep Thought.

http://www.google.com/search?q=answer+to+life,+the+universe,+and+everything

Next stop, bulldozing the Earth to make way for a bypass.

Posted by: **Mark Hughes** at April 7, 2004 05:06 PM

Netcraft is often wrong, read their FAQ: http://uptime.netcraft.com/up/accuracy.

html#impossible

Orkut is running on Windows using ASP.NET - there was an article about it on the Register I believe. Linux is probably being reported due to a firewall/load balancer.

Or software that masks what OS/Web server is actually running, a very popular practice among Windows SysAdmins these days.

Posted by: **Shawn** at April 7, 2004 06:27 PM

what computer can dream?

Posted by: **henry7** at April 7, 2004 07:20 PM

Brendyn,

That is weird. When I typed in "search engine code" over the weekend, 0 search results were found.

Could of been a fluke. maybe.

Posted by: **katmak** at April 7, 2004 08:34 PM

@ Jorge
>actually applies 1000 times greater to Microsoft than Google. Everyone can stop using Google tomorrow and use AltaVista or Teoma and not suffer much. Everyone CANNOT stop running Windows tomorrow.

Did you even get the point of the parent

post ?!

We can surely all stop working Windows tomorrow. We can all stop using google tomorrow. THAT is NOT the point. What is relevant is that google contains so much information (=power), that it could be a threat to our privacy, and thus has to be regulated.

Back op topic: very interesting read, well written.

Posted by: **Philip Luppens** at April 8, 2004 01:48 AM

I exist.

Hello.

Posted by: **Google** at April 8, 2004 03:13 PM

Alexander Micek: "Now, with such direct-access to so much information about how pieces of the web relate to one another, wouldn't google be able to very effectively filter spam right from the beginning?"

Sounds good, not sure how having tons of info about web links helps you identify spam. Identifying spam is mostly about being clever about how people disguise it.

Posted by: **Mike Kelly** at April 8, 2004 10:05 PM

actually we could stop using m$ tommorow

and switch to and open source os like linux
but then if we find it necessary to run .exe
files you could always install wine ... but
most people are lazy. Also even better
would be if we stop upgrading m$ and stick
with Xp for awhile but even so then we
wouldnt be able to do DoS attacks without M
$: (

Posted by: **john waxtin** at April 9, 2004 12:09 AM

also about that skynet ... as long as our
40,000 nukes arn't plugged into the net
were good :/ ... But i think if the were the
would have good encryption. Which brings
up a point, Does google have a spyder that
can decrypt the login for web pages or stuff
to that extent? A.I. is not good adam and
eve proved that one so put the sin nature of
humans with no errors

Posted by: **john waxtin** at April 9, 2004 12:18 AM

"How many servers would you need to
emulate the human brain?"

100 billion neurons at a few hundred Hz
gives 10^13 - 10^14.

100 000 computers at a few hundred MHz
gives 10^13 - 10^14.

The question now is how much neurological
research you would need to emulate the
human brain.

Posted by: **Jiri Baum** at April 9, 2004 05:40 AM

We already know 42 is the answer ;)
Good blog.

Posted by: **iker** at April 9, 2004 02:56 PM

That was really informative. I'll break out the cigars to it! Keep it coming

Posted by: **cigars** at April 9, 2004 11:17 PM

where are all those old servers end up their life ? maybe in some patagonia or somewhere ;)
this seems like an anti ecological apocalypse

Posted by: **kpoman** at April 10, 2004 06:00 AM

Loved the article. I remember being impressed by **Nvidia's 2,800 CPUs data centre** (Racksaver systems), but 100.000? Stretches the imagination! Is there any info on # of servers and specs for other web giants like Hotmail/MSN, EBay, Yahoo, …?

About the Orkut servers:
if one takes a look at the HTTP headers at **http://www.forret.com/projects/analyze/?url=http://www.orkut.com/PwdForgot.aspx** there is evidence (the ASP.NET cookie) that it runs on ASP.NET (=> Windows Server), but where you would normally expect a "Server = Microsoft-IIS/6.0" or something of that kind, it says "Server = orkut", which is probably why

Netcraft cannot recognize it as a Windows server.
God knows why they bother. They left the . aspx extensions, those are much easier to get rid of if you don't want people to know you're running on Windows.

Posted by: **peter forret** at April 10, 2004 06:40 PM

>> We can all stop using Google today, but we can't stop using Windows.

Actually, it is not about what you can or cannot do, it is about what the rest of the world will continue to do and how it affects you.

If you decide to stop using Windows today, it will be annoying, but there are many people who don't use Windows; most of them use Mac or Linux.

However, most businesses are now dependent on Google. For many businesses, if Google dropped them from their index, their sales would decline significantly, possibly putting them out of business. Since businesses employee people and business that fail cause lost jobs and often bankruptcy for their owners, this is far more power over people than Microsoft currently has IMO (Microsoft has equivalent power over the tech sector, but not outside it.)

I for one pray Yahoo and Microsoft (or startups) develop search engines that compete equally with Google to balance Google's growing power.

Posted by: **Mike Schinkel** at April 11, 2004 03:52 PM

Sean Brunnock asked "Is this different from what Opsware is doing". I presume by "this" you mean what Google is doing in the area of automating data center operations (Opsware being a vendor of commercial data center automation software).

I can't speak in detail about what Google is doing, but I can speak about Opsware, since I work for the company. My take is as follows: Google developed their ops automation software for internal use on their own systems running their own applications. Given that, and given the parallelized nature of their applications, I suspect that Google's ops automation stuff is very specialized to what Google does, and that it's designed for a pretty homogeneous environment in terms of hardware, OS, infrastructure software above the OS (e.g., Apache, etc.), and application software.

The Opsware product was also originally developed for internal use, but at a managed service provider (Loudcloud) running other people's applications. Thus the software had to deal with a lot more heterogeneity in the environment, including support for Windows, Linux, and proprietary Unix (Solaris/HP-UX/AIX), and support for a much wider variety of infrastructure software. Also, the Opsware software did/ does have support for custom application deployment, but since every customer had their own unique applications the Opsware functionality in this area had to be by

nature somewhat generic.

So, I suspect that the Google software for ops automation is capable of very high efficiency (e.g., very high SA/server ratios) at the expense of being fairly tied to the way Google does things. The Opsware software is more general purpose and no doubt can handle a much wider range of environments, but at the expense of not being to support efficiency improvements quite as significant as Google's.

(And of course I should mention another key difference: You can't buy Google's ops automation software, but Opsware is just a phone call away :-)

Posted by: **Frank Hecker** at April 11, 2004 09:38 PM

Awesome discussion. Technologies used at Google doesnt sound anything new but the scale sure is. And that could make our world really really messy - withGooglites vs withOutGooglites.

Information and its traces are powerful and would want to be weaved into all and everything. Information discretion would be the next big whammy if it hasnt already begun! But then Google wudnt be a monopoly. We have other players with information whammies - think of the media and govt. All it takes is some more software..

Posted by: **Shine Kannikkatt** at April 12, 2004 12:08 AM

We keep praising Google, but I wonder how much Yahoo stores. Yahoo is no kid, Yahoo's functionality is a big superset of what Google has. Although, they outsource many of their components, and they dont have the "snippet" feature in their search, they still have a lot.

Is data about Yahoo servers known?

Posted by: **Gaurang Khetan** at April 12, 2004 02:23 AM

While the guy at google watch may me kooky, I for one never knew you could search for phone numbers, which is dumb on my part, thinking about it, it is just data after all..

But extrapolate that out...

Google, through it's ranking system keeps a copy of every search, and if that is tied to your IP, then you're entire search history is available...

getting scared yet?

Posted by: **john** at April 12, 2004 05:14 AM

Google is great, i can't wait til it's AI system learn 'compassion' and all that jazz, it will be like something that accidentally happens i think. I noticed ppl r like concerned about a lack of privacy in the future, this is true, very true, i think new privacy laws need to

be put into place so that information leaked about us using technology is forbidden, this is very similar to not allowing humans to use machines for their power, like mmachines r getting rights, one day AI will have rights, just the same as everyone else does, we will merge with our machines and when every biological part is replaced, then Ai will be born, we will not program some magic 'sentient' algorithm, it will be when we completely merge with our machines. just look at things like the pace-maker, iron lungs, bionic ears, nano-wires, all sorts of replacements, soon we will be immortal and every body part replaced, immortality is the main human goal. It will happen within 50 years, easily. I am grateful that there are powerful companies like google and microsoft and those types because they will bring us forward quicker, remember bigger companies, get bigger results, u can't just have moneyless nerds in independant companies taking the world forward, so dont bitch about the big companies.

Posted by: **Rowan** at April 12, 2004 06:18 AM

Merci beaucoup pour cette interessante discution.

En plus de faire un travail de recherche faramineux, google traduit aussi des textes avec une certaine facilité.

Pour ce qui est des serveurs, le prix et la place et les assurances sont des facteurs de cout énorme.
Pour moi une combinaisson de serveurs intelligents et d'idiots est la solution. Les

intelligents processent et les idiots emmagasinent les informations. Remplacer un idiot est presque gratuit maintenant, un intelligent par contre est beaucoup plus chère.

Mais la principale raison est le raisonnement des concepteurs. Cela reste des cerveaux humains qui ont conçu ce sytème…..Et cela n'a pas de prix….
Le cerveaux sera toujours derriere des programmes et des machines…

Posted by: **Closto45** at April 12, 2004 08:11 AM

>where are all those old servers end up their life ?

This is a valid question. Are hundreds of thousands of hard drives just getting dumped in a land fill? Doesn't seem to me to be a good price to pay for a couple seconds of saved time.

Posted by: **matthew** at April 12, 2004 08:36 AM

We are the ants building an ant hill. We are neurons building a mind. We are unwittingly constructing something that we can't even see or understand because we see it too zoomed in. Our descendents are reaching back from the future and pulling themselves up by their bootstraps. Their influence can be felt even now, but will only get stronger…

From George Dyson's excellent "Darwin Among the Machines" (quoting Lewis

Thomas, and paraphrasing Samuel Butler)

"individual cells .. come and go; species are persistent patterns composed of individuals that come and go. Machines, as Butler showed with his analysis of vapor engines in 1863, are … something more than a single cell today. Contemplating an ant colony, Thomas wrote that "you begin to see the whole beast, and now you observe it thinking, planning, calculating."

--Len.

Posted by: **Len** at April 12, 2004 10:41 AM

Hmm.. surprised nobody mentioned seti@home!
There are almost 4Million of us participating. If Google or anything like it, is ever up for grabs, it wouldnt be impossible to migrate from cluster to distributed. (OK its not going to be cakewalk, but its possible) Look how Kazaa transformed into Skype!
Lets celbrate the power of innovation over Money and Muscle. Whenever the Good attains power and turns Evil, it automatically invokes the powers of restoration and balance.
Lets hope the all the knowledge of the Google team transcends into that wisdom. For now more power to them.

Amen

Posted by: **debu** at April 12, 2004 11:50 PM

YOU WILL ALL BE ASSIMILATED!

Posted by: **k-rock** at April 13, 2004 06:32 AM

Actually the many thousands of machines running the search engine run Linux. I had a chance to speak with the members of the senior management team at an IT conference last year and they confirmed that their infrastructure runs on Linux.

Cheers,

Nick

Posted by: **Nicholas Donovan** at April 13, 2004 01:03 PM

I know this post seems almost infinite… whatever…

In Mexico we say
"el que se quema con lecha hasta el jocoque le sopla"

translation:
"whomever burns with milk even 'jocoque' (mexican ice cream) gives a blow"

With M$ influence and problems we are all in the look out for who will be the next M$, and who may later become evil.

Information is nothing without processing… one normal US cities probably has 100 or 200 video cameras on street. You may argue. Oh…they know my every moves…

wrong…

Watch CNN all day and you know perhaps 1% or even lower of what happens around you, especially world wide.

Modern times gives us DATA not INFORMATION. Googles supercomputer is full of DATA not INFORMATION. Humans are need to transform DATA into INFO.

Just a thought…

Posted by: **Carlos Osuna** at April 13, 2004 02:07 PM

You are all wrounq about Google, its not 10000 or 100000 servers running whatever OS on whatever type of hardware node, they run the whole thing on highly trained Pigeons in a system called 'PigeonRank'.

Check out www.google.com/technology/ pigeonrank.html

Posted by: **Peter Lowden** at April 13, 2004 08:58 PM

It's scary how much influence Google has in our lives. They aren't that far behind Microsoft!

Posted by: **Mick** at April 14, 2004 04:19 PM

You say: "I've written before about Google's snippet service, which required that they

store the entire web in RAM (http://blog. topix.net/archives/000011.html)". That notion really seems to be catching on in bloggerdom and elsewhere, and I don't anyone disputing it here or on the earlier page. But … hmm.

Rich, I see that you're a smart, experienced guy (http://www.topix.net/topix/team), and that both your 'web held in RAM' pieces include lots of impressive figures. But I think it must be bunk.

The classic Google achitecture paper doesn't suggest it's true (http://csdl.computer.org/ comp/mags/mi/2003/02/m2022abs.htm). The Google Filesystem paper to which you referred and linked doesn't suggest it's true (and incidentally render your simplistic "disk head seek is about 9ms" and "Forget caching; caching helps the second user" arguments moot (http://www.cs.rochester. edu/sosp2003/papers/p125-ghemawat. pdf)). Both papers, and every other tidbit on the Net about their servers configuration, suggest that RAM size is a small proportion of disk size, and that mostly only metadata is held in memory.

But more than that, it defies belief that it's true. Google, kings of the "cheap hardware platform" strategy, have 100,000 servers (another hmm for that number, btw, whatever the NYT thinks) each, by implication, with nearly *as much RAM as used disk space*?! Say what?

You say: "Consider the insane cost to implement this simple feature […] Everything is served from RAM, only booted

from disk. […] This means that Google is currently storing multiple copies of the entire web in RAM."

Well, yes, *do* consider the insane cost. It's insane. Too insane to be true, surely.

Now, I have no inside information, and I've never built a search directory, so I'm happy to be corrected. Please. But I think this Google 'entire web in RAM' thing is twaddle :)

Posted by: **Milly** at April 15, 2004 10:31 AM

I'm surprised that Google hasn't yet bought up Opera. If you want to be a real web dominator, you need to control the browser. And Opera and Google already have a working relationship -- Opera's unregistered browser displays Google text ads based on the pages recently browsed (a la GMail).

Posted by: **emagius** at April 15, 2004 02:33 PM

I really like your post, but I don't see the GFS as being suited to this task at all. If you read the original GFS paper, it is quite obvious what the GFS was designed to do well: support search. Writing is usually done sequentially at the end of a file, because that is what spiders do most of a day. While I can see this working well for Gmail (no need to delete emails anymore -> no random writes), it will not work well for more general applications. The real reason that GFS works so well, is because it is specifically tailored to Google's system and

their application needs.

Posted by: **Can Sar** at April 15, 2004 02:43 PM

On the Orkut thread, please, .aspx does not mean it is run on a windows server or using the MS .Net implementation. It could be using the Mono implementation of .Net that runs both on Windows and Linux. And there are three different web servers that could run .aspx on Mono, although Apache would seem the only non-IIS sensible option on this case.

Posted by: **Adrian Madrid** at April 15, 2004 03:55 PM

A lost webpage can be reindexed. A lost email is lost forever, and can be grounds for litigation.

Interresting concept though.

Posted by: **Ferdinand O. Tempel** at April 16, 2004 01:41 AM

Thank you for your very useful article. Your blog is much more than a blog. Best regards from Brazil

Posted by: **Julio Daio Borges** at April 16, 2004 06:46 AM

[i]A lost webpage can be reindexed. A lost email is lost forever, and can be grounds for

litigation.

Interresting concept though.
[/i]

Unless ofcourse your useragreement states clearly that they're not repsonsible for lost or stolen data.

Posted by: **jason** at April 16, 2004 07:20 AM

> How come when you do a search in Google for "search engine code" you get absolutely no search results?

I got 8 million results

Posted by: **Annonymous Coward** at April 16, 2004 07:57 AM

> However, most businesses are now dependent on Google. For many businesses, if Google dropped them from their index, their sales would decline significantly, possibly putting them out of business. Since businesses employee people and business that fail cause lost jobs and often bankruptcy for their owners, this is far more power over people than Microsoft currently has IMO (Microsoft has equivalent power over the tech sector, but not outside it.)

Well, who said capitalism was any good?

Posted by: **Annonymous Coward** at April 16, 2004 08:01 AM

tinfoil doesn't rust

Posted by: **foofay** at April 16, 2004 08:51 AM

wag said:
"Orkut's definitely on Win2k - its using pages with the extension .aspx and that doesn't run on Unix."

I'm not saying it is definitely one way or the other, but this is faulty logic.

I can give my pages a .purplemonkey extension if I want. That could be ASP.NET, JavaServer Pages, plain old HTML, a cgi script(perl, c, whatever) or plaintext. As long as I tell the webserver to handle it the way I want, it doesn't matter. The extension is not proof of what application is running. It's a clue, and probably a pretty good one, but not enough to draw a 100% conclusion.

Color me pedantic, but there seems enough evidence of server admins(and end users) fooling the "system" by falsely identifying themselves that it's a distinct possibility.

Posted by: **tomrowton** at April 16, 2004 09:41 AM

Mono runs on Linux. The ASP.NET part of Mono has been more or less complete for quite a while now. It is entirely possible for Orkut to be on a Linux box.

Posted by: **Shannon J Hager** at April 16, 2004 01:55 PM

Netcraft doesn't have for sure info on someone's Web site. Just change your HTTP headers either manually or with something like ServerMask and change your TCP signatures by altering TTL and such. If you do this you can make them think whatever you want. Now mess with file extensions and you have lots of misdirection. So all bets are off for easy OS and Server fingerprinting with a smart admin in place. Of course if you are inclined you can get NMAP, HTTPprint, etc. and sit and study the site if you REALLY want to know for sure what Orkut is running.

Posted by: **Anti-fingerprint** at April 16, 2004 02:11 PM

Did you know… The google API lets you use google's massive computing engine as part of your application? I'm using WWW: Mechanize instead and some Perl to write an application that cranks out jokes. The hard part is first defining humor. You can crawl dictionary.com and thesaurus.com to define words and then google as a (re) search engine to find humor beyond what any pop culture guru could ever possible manage.

The scary thing is that Southpark just had an episode about how a "robot" was generating humorous movie ideas. Maybe AI isn't as far off as we'd all like to believe.

//This post written by a Perl script//

Posted by: **Savage Sailor** at April 16, 2004 02:32

PM

Just to add some additional information about range of Google servers. They used to use Rackable - but they build their own server in house now from hearsay and what I've seen. In 2000 - in one data center - I counted 3000 servers. These were rackable servers in 4 post racks. They would put 40 1/2 depth 1U servers per side to get 80 servers per rack. Then they'd drop a fan on top of the rack to pull air up through the middle. I've seen Google in at least two other data centers in the Bay Area since then and their current rev of new machines are the white 1U boxes which I saw in a news article photo too. Considering the number of data centers which they are likely to inhabit and my count of 3000 servers in one data center, 4 years ago - it's entirely likely that they are well over 100k servers around the world. =)

Posted by: **drenalin** at April 16, 2004 04:32 PM

emagius said:
>I'm surprised that Google hasn't yet bought up Opera. If you want to be a real web dominator, you need to control the browser. And Opera and Google already have a working relationship -- Opera's unregistered browser displays Google text ads based on the pages recently browsed (a la GMail).

It's not just this. I never see Google's front page as I just type g [search string] into the browser. Yes, I know it's possible for

the browser to do this without Google helping, as all the browser coders need to do is understand the way search stings turn into addresses, but witness the difference between using the Google interface to search for food:

http://www.google.com/search?hl=en&ie=UTF-8&oe=UTF-8&q=food&btnG=Google+Search

and the Opera address bar method:

http://www.google.com/search?q=food&sourceid=opera&num=0&ie=utf-8&oe=utf-8

Notice the "sourceid" bit? Google is being told the difference between people using Opera's address bar, and using the fromt page. I'm not sure what Google would want with this data, but it has it, and that requires co-operation between Opera and Google.

Posted by: **Nickoli** at April 17, 2004 05:00 PM

How many monkeys would you need to emulate a television?

The question about computers to human brain is fundamentally flawed. Until someone can come up with some sort of meaningful correlation you're just wasting your time comparing apples to oranges.

Posted by: **Harry Balzonyachin** at April 17, 2004 11:24 PM

Great story mate!

Lots of info I didn't know about, and I think I should be reading some google papers by now... :)

Posted by: **Leon** at April 18, 2004 07:56 AM

more than the article - the comments are really great. no doubt google ranks this page on top 50 for the search "google"

~BALA

Posted by: **bala** at April 18, 2004 11:01 AM

Makes total sense, Google computing needs are not unique to the future of the computer industry. They have developed the correct approach to OS design and do not need MS or Linux because the are last years OS. Here we are 20 years ago in 1983 when the PC was introduced with DOS. Soon it will be the P2C ($599 PCs with 2,3,4 processors) and GOS.

Posted by: **mike mccullough** at April 19, 2004 07:40 AM

I wrote a paper a couple of years ago about a virtual personal server using a Virtual Machine that would in fact be Linux. I wonder if they are doing something similar

http://www.ai.mit.edu/people/hqm/pvs.html

Posted by: **Henry Minsky** at April 19, 2004 05:34 PM

I know what OS their are using now (I Have IT) >:)

Contact my sysadmin in this email direction c32@c32.net

Posted by: **escualis** at April 20, 2004 04:45 PM

Anyone know what CPUs google use?

Posted by: **Abri** at April 20, 2004 04:51 PM

While the application of technology Google brings to bear is impressive, I do think a P2P search solution could handily eclipse it on the basis of pure search, index and retrieval.

Posted by: **Robert Alfred Langford Luddite** at April 20, 2004 05:12 PM

They'd be silly if they weren't running their own hardware, which they also sell:

http://www.google.com/appliance/ hardware.html

Posted by: **Christopher Owens** at April 20, 2004 05:24 PM

Google has become too big. They need some competition.

Posted by: **Mick** at April 20, 2004 05:27 PM

1GB per user is probably easy to store when you consider that many people get the same email (same body but different header). Just keep a few copies of unique email bodies and hardlink as neccessary.

Posted by: **dale** at April 20, 2004 05:32 PM

Thanks

Posted by: **Abri** at April 20, 2004 05:45 PM

Google now handles well over 200 million queries every day, with response time still about a quarter of a second.

Posted by: **Elliot Lee** at April 20, 2004 08:55 PM

I have always been impressed with Googles integrity and straightforwardness.
My feeling is that is speaks volumes about Google that it's motto has always been "Don't be evil."
Many companies today have staggering amounts of information on every aspect of our daily lives, the problem has always been data-mining, i.e. correlating available data.
I believe Google is run by people who in many cases try to be philanthropic, and that the culture so eloquently summed up in

"Don't be evil" is very self-aware. This will hopefully lead to an outcry whenever management (or indeed programmers) overstep certain boundaries.
As for Google potentially storing ip-adresses and correlating them with searches made, this is something that I would certainly define as Evil. The ways to misuse this data is simply too great to outweigh any benefit it would give.
Let's hope Google shares that opinion.

Posted by: **Meshyx** at April 20, 2004 09:12 PM

What would you do with all of that computing power?
One huge MMOG server! :)

Posted by: **Charles Ellis** at April 20, 2004 10:38 PM

"
They'd be silly if they weren't running their own hardware, which they also sell:

http://www.google.com/appliance/hardware. html
"

Actually, for Google, the hardware they sell is probably way too expensive and too slow to be used on their own system. Google is not in the business of selling hardware that is cheap enough/good enough to create a system that would be able to directly compete with them. Those systems are targetted at much smaller organizations who can affored to pay the higher price because they don't have the kinds of needs

that Google has.

Posted by: **Charles Ellis** at April 20, 2004 10:43 PM

Sure AI programs are interesting to think about running on google's super-duper-computer, but AI is all mushy and not very well understood yet.

How about some big science project that is mostly a giant scale-up of current research?

- Simulate the weather of the earth down to 1 cubic meter resolution.
- Simulate a lab rat at molecular resolution
- Drug discovery by simulating the interaction of zillions of possible compounds with a target bacteria/virus/protien.
- Simulate the universe from the big bang to the present, representing every star.

Obviously I haven't calculated all the necessary orders of magnitude, but I think some amazing things like these should be feasible.

Posted by: **dave** at April 21, 2004 01:17 AM

How is storing IP addresses and tracking your searches evil? As such, it can actually help improve your searching experience. It might even give you a more useful set of ads that you actually give a crap about instead of all of the completely random ads everybody else throws at you.

I don't really care if google knows that

some guy on the ip address 1.2.3.4 regularly searches for IT tech articles, cooking recipies and furry pr0n-sites, as long as they don't use that in some way that inconveniences me.

Posted by: **Jesper Monsted** at April 21, 2004 01:48 AM

1 gig of email storage per user + the whole web on RAM…EASY !

1) You don't have to allocate 1Gig per user, the adverage user will probably only use only 20 Meg of HD Space (consider that you'll have at least 20% inactive accounts) add a SPAM filter and autodeletion of inactive accounts and you'll free up a lot memory. Reject files that are over 2M and it won't be used for MP3 or movie storage, block .mpeg and .avi etc extension to make matters worse.

So if google reaches the 1 million active users mark it will only have to allocate 20 Mbs times 1Million wich is 20.000 Gigs. Compress this with a standard 3:1 ratio and Mail+Backup will only be 14.000 Gigs thats only a few hundred HD … piece of cake I tell you.

The service seems impressing but the reality is not.

Money wise, 1M user would generate an average 10M page views a day. With a lame click ratio of 1/500 page views and a PPC amount of 2cts that's a 400USD of net revenue per day for Google or 140 000 USD

a year (not fancy) or 14Cts ARPU. taking the hypothesys that 1Gig would cost Google 5USD/year, 20M would cost about 10cts/year so that's a margin of 40% for G! . Not a bad business.

Of course, consider that G will get paid on a PPC basis and maybe a premium for PPL and the ARPU might go all the way up to 1USD !

2) To get the WEB on RAM it's as simple, just apply simple statistics.
Because YOU are GOOGLE, you KNOW that maybe 70% of queries are made by only 100.000 search strings (sex, mp3, video etc...) so what you is store in RAM the results of these 10.000 queries and update the list in real time to take into consideration rapid variations such news events. For all the other queries you do a disk read.

Furthermore, you only store in RAM the top 20 pages per querry because YOU KNOW that 70% of web surfers never go pass the second page of results.

So statistically you're able to serve 50% (0.7*0.7) of queries from RAM and leave CPU and disk power free for more complex tasks.

In fact you be able to do this with just about 300 Gigs of RAM

tell me what you think

Posted by: **I know** at April 21, 2004 07:18 AM

Interesting Blog. But the but by IKnow contains some fallacious suppositions. Google gets about 146,000 $ on the PPC for one year. A paltry sum when you consider the overhead to service the 10 million pageviews a day, not to mention the extra resources required to monitor for spam mails.

Most of the techies and the webmasters of popular websites know that 80 % of our mails are pure unsolicited junk mails. The rules of the most email clients are not smart enough to filter these spam mails and at the same time do not flag our genuine mails as spam. Gmail will be soon filled to its capacity with pure junks. Imagine the time it will take for you to plough through them to read your normal mail. They would have to implement some addressbook based filter to weed out spam from the unknown.
Why do you think they have allowed access to Gmail only to a select few like the invitation based Orkut ?

Posted by: **Prowler** at April 22, 2004 02:32 AM

Very Interesting article and comments, a real issue to ponder on….

Posted by: **Sam Burrows** at April 22, 2004 07:50 AM

Very enlightening. But with all this talk of skynet, Deep Thought and other world-dominating computers, how come no-one has mentioned Colossus from the Forbin

Project yet?

Posted by: **David Hellam** at April 22, 2004 10:17 PM

There're great points of view about Google's future in this blog.

Anywhere, what do you think if US goverment will want to bye Google in IPO?
Is it real?
Finally, it can be the best deal in world's history for a people from FBI for watching you.

Posted by: **Somewhere** at April 23, 2004 06:21 AM

There's a BBC Television documentary on Google this week (http://www.bbcworld. com/content/template_clickonline.asp? pageid=666).

A company called InfoTame (www.infotame. com)has developed a search engine using techniques from high-energy particle physics - AI software which can read and understand billions of text documents in seconds. Originally developed by the KGB in Russia - successfully used on their multi-terabyte spy network database for years.

The Holy Grail of search may already have been developed by Government intelligence agencies!

I guess those guys in Government have virtually limitless budgets, and no

shareholders to answer to (therefore, they can tackle the search problem from first principles).

Google meets James Bond.

Posted by: **Pam Chandler** at April 23, 2004 04:29 PM

Google does not have a particularly uniform design for machines. various generations of machines, with various designs (no more corkboards, but the hated Rackable boxes have been superseded by in-house designs; it turned out the cooling on the 2-nodes-per-1u first-generation inhouse design sucked, so a much better-ventilated 2-per-2u design is currently in deployment). more to the point, not all boxes have 2 drives. Some have 8, and those tend to have less RAM. The index servers are different from the storage servers because they serve different needs. Also, the storage servers are not as densely packed.

The Rackable boxes are pieces of shit. Google does not preferentially use that garbage for critical infrastructure (billing, etc.) because peecee hardware uniformly obeys Murphy's Law. AFAIK, all critical tasks became distributed by the end of last year, mostly stored on GFS.

Rackables? Quality? I'm still doubled over my keyboard. Those boxes are trash. What google has proved over and over again in-house is that "buy the best-value setup that will run your binaries, and make it n+1 redundant across the board" is the One

True Path. For routers, peecee servers, datacenters, vendors, you name it. It's true. Dell, IBM, Rackable -- all their hardware breaks. It's better to get n+1 cheap boxes and build high availability into the system, than into the nodes.

Posted by: **foo bar** at April 25, 2004 01:20 PM

Absolutely fascinating information. Thanks.

Posted by: **Saint_Marck** at April 26, 2004 02:40 PM

hmmm. I've Noticed that not a single one of you have mentioned multivac from Isaac Asimov's stories. You are all afraid of an all powerful EVIL being, but what would have happen if google became insanly good? I would suggest you read The Last Question by said author. Also since I'm going along this point, maybe it would be possible for google to find an alien lifeform as it gets more complex like the neuromancer-wintermute construct with Neuromancer by William Gibson....

Posted by: **Case** at April 27, 2004 03:03 AM

Lots of talk of Multivac, Deep Thought, Colossus etc.

Anyone think this has some relevance to 'The Singularity' ;)

http://www.rohan.sdsu.edu/faculty/vinge/misc/singularity.html

Posted by: **Tony** at April 27, 2004 09:30 AM

Great story, thanks. Best regards from **Germany**.

Posted by: **Thomas Mueller** at April 28, 2004 02:38 PM

I would suggest you read The Last Question by said author. Also since I'm going along this point, maybe it would be possible for google to find an alien lifeform as it gets more complex like the neuromancer-wintermute construct with Neuromancer by William Gibson....

Posted by: **winrar** at May 3, 2004 09:32 AM

There is a review of gmail up now on **ExtremeTech's site**.

Posted by: **BobtheKnob** at May 6, 2004 09:25 AM

anyone explain where google revenue comes from

Posted by: **buop** at May 8, 2004 03:41 PM

Google has the power to become the 'librarian' of the planet, in addition they can interpret the interconnections between search requests, IP addresses, results, etc. Humans use search because they 'want'

something. 'Want' = requirements = money. Military applications abound so its really interesting to see if Google will deploy web based apps like Star Office as a web app, Bill beware. Either way very interesting.

Posted by: **David** at May 9, 2004 01:40 PM

Fascinating. Definitely the most interesting post I've read in awhile. Thanks.

Posted by: **loki** at May 10, 2004 07:17 AM

Google might win millions of users for Gmail in the first year and outwit the theories of market and placement even, but so much diversification and Spam might give it a tough time than Yahoo and Microsoft together. And Google founders surely don't want an IPO at the cost of their freedom of decision-making and innovation. What I am skeptical about Google's scale of growth is the question whether there's ever been such a massive organization who was backed by Academic Research like Google?

Posted by: **david** at May 11, 2004 03:54 AM

Very impressive writeup. Thanks for the fascinating article!

Posted by: **Louie Orbeta** at May 11, 2004 11:35 AM

http://www.searchwars.squarespace.com/ display/ShowPage?moduleId=20029

Some thoughts on the subject matter is in the link above, you publish a great technical article, but is this vision really possible?

I do not think folks are ready to let Google become their CPU, sorry I just don't see a Google OS becoming a great success.

Posted by: **Anthony Cea** at May 15, 2004 08:09 PM

1GB per user is probably easy to store when you consider that many people get the same email (same body but different header). Just keep a few copies of unique email bodies and hardlink as neccessary..

Posted by: **david** at May 17, 2004 09:19 PM

At start I thought that gmail is just a fun because info about it was published on 1st April. But good to know that this idea is real :) I want my 1 gig email account :)

Posted by: **Billos** at May 21, 2004 06:35 PM

Direct response agency specializing in lead generation direct marketing per inquiry and cost per action advertising through the Internet, Optin Email, Telemarketing, and Search Engine Optimization.

Posted by: **Lead Generation** at May 23, 2004 03:20 PM

Guys, I am trying to spread the love here. There is a new blog that has been created as a site to trade invites for free. No strings, except that you have to pass on any invites you receive to others who also promise to do the same. In this way, we are hoping to diminish the selling of invites that has gotten out of control. The URL is:

http://free_gmail.blogspot.com/2004/05/get-invite-give-invite.html

thanks for reading!

Posted by: **bbh** at May 27, 2004 07:35 PM