S. NORFOLK

Left: the data centre at the Wellcome Trust Sanger Institute in Cambridge, UK, under development.

WELCOME TO THE PETACENTRE

What does it take to store bytes by the tens of thousands of trillions? **Cory Doctorow** meets the people and machines for which it's all in a day's work.

Ten seconds after I stepped into the roar of the data centre at the UK Wellcome Trust Sanger Institute, in rural Cambridgeshire, my video camera croaked: CARD FULL. Impossible. That morning, I'd tossed a handful of thumbnail-sized 32-GB memory cards into my pocket, each one good for a couple of hours' worth of high-definition video. Yet this one had filled up in seconds.

I fumbled with my camera while Phil Butcher, the Sanger Institute's head of information technology (IT), politely waited, grinning in the shower of cold air washing down from the air conditioning. It took only a couple of embarrassing seconds to troubleshoot: I'd somehow mixed an old 32-megabyte card in with the 32-gigabyte cards. The 32-MB card is only a couple of years old; when I bought it, it probably cost more than the 32-GB cards do today. But it holds one one-thousandth of the data.

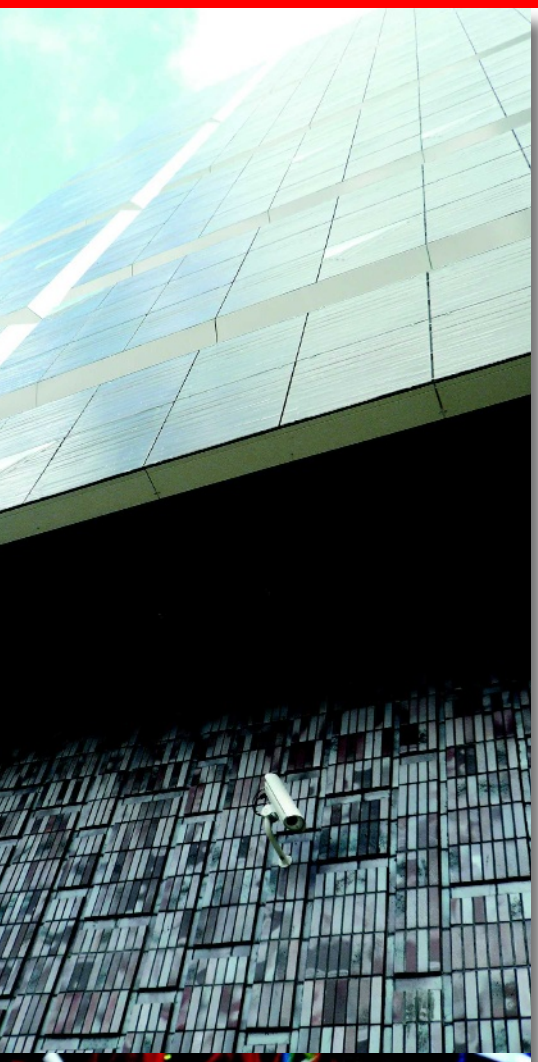
That, in coincidental microcosm, is the story I'm here for: the relentless march from kilo to mega to giga to tera to peta to exa to zetta to yotta. The mad, inconceivable growth of computer performance and data storage is changing science, knowledge, surveillance, freedom, literacy, the arts — everything that can be represented as data, or built on those representations. And in doing so it is putting endless strain on the people and machines that store the exponentially growing wealth of data involved. I've set out to see how the system administrators, or sysadmins, at some of the biggest scientific data centres take that strain — and to get a sense of how it feels to work with some of the biggest, coolest IT toys on the planet.

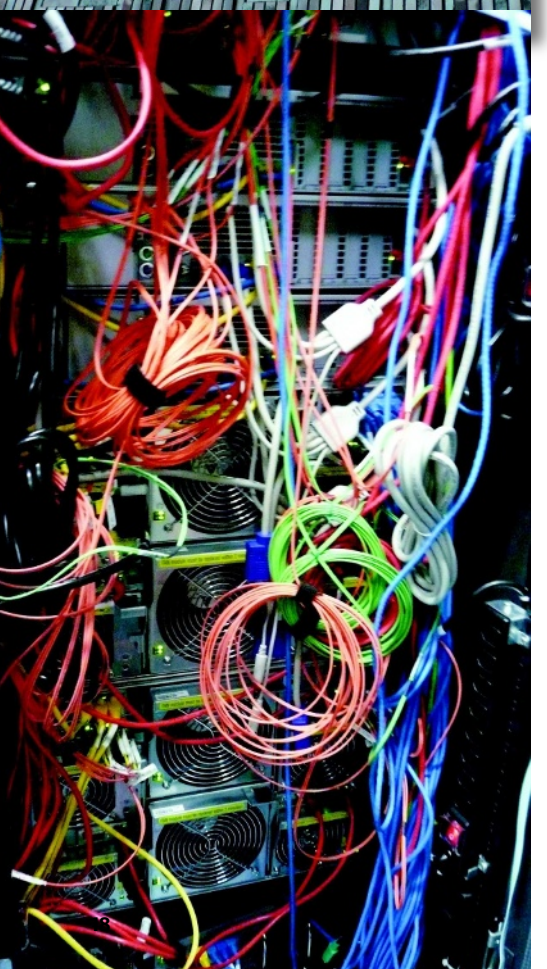
At this scale, memory has costs. It costs money — 168 million Swiss francs (US$150 million) for data management at the new Large Hadron Collider (LHC) at CERN, the European particle-physics lab near Geneva. And it also has costs that are more physical. Every watt that you put into retrieving data and calculating with them comes out in heat, whether it be on a desktop or in a data centre; in the United States, the energy used by computers has more than doubled since 2000. Once you're conducting petacalculations on petabytes, you're into petaheat territory. Two floors of the Sanger data centre are devoted to cooling. The top one houses the current cooling system. The one below sits waiting for the day that the centre needs to double its cooling capacity. Both are sheathed in dramatic blue glass; the scientists call the building the Ice Cube.

Blank slate

The fallow cooling floor is matched in the compute centre below (these people all use 'compute' as an adjective). When Butcher was tasked with building the Sanger's data farm he decided to implement a sort of crop rotation. A quarter of the data centre — 250 square metres — is empty, waiting for the day when the centre needs to upgrade to an entirely new generation of machines. When that day comes, Butcher and his team will set up in that empty space the yet-to-be-specified systems for power, cooling and the rest of it. Once the new centre is up, they'll be able to shift operations from the obsolete old centre in sections, dismantling and rebuilding without a service interruption, leaving a new patch of the floor fallow — in anticipation of doing it all again in a distressingly short space of time.

The first rotation may come soon. Sequencing at the Sanger, and elsewhere, is getting faster at a dizzying pace — a pace made possible by the data storage facilities that are inflating to ever greater sizes. Take the human genome: the fact that there is now a reference genome sitting in digital storage brings a new generation of sequencing hardware into its own. The crib that the reference genome provides makes the task of adding together the tens of millions of short samples those machines produce a tractable one. It is





C. DOCTOROW

what makes the 1000 Genomes Project, which the Sanger is undertaking in concert with the Beijing Genomics Institute in China and the US National Human Genome Research Institute, possible — and with it the project's extraordinary aim of identifying every gene-variant present in at least 1% of Earth's population.

As data pour off the Sanger's new Solexa sequencers, Butcher — a trim, bantam grey-haired engineer with twinkling eyes and laugh-lines — has to see to it that they have somewhere to go. A two-hour Solexa run produces a gigantic amount of information: 320 TB, according to Tony Cox, head of sequencing informatics, a figure he's mentioned to journalists in the past (a print-out on his office door reads: "'Oh shit, that's 320 TB!' — Tony Cox, *The Guardian*, 28 February 2008"). The 1000 Genome Project needs to make use of storage and computing capacity at a (currently) impossible density. Luckily for Butcher, 'impossible' is a time-bound notion — if you don't like the compute reality, just wait around a little while and another will be along shortly. His storage density is doubling every year; the 500-GB hard-drives spinning away in his storage array are being phased out by Seagate of Scotts Valley, California, the company that makes them, in favour of a terabyte model.

"The PetaBoxes have the elegant, explosive compactness of plutonium."

Finding a place for the data to go is only the beginning. Butcher also has to make sure they can get back out. The Sanger has a commitment to serving as an open-computing facility for the worldwide research community. So it faces what you could call the Google problem: an unpredictable and capricious world that might decide at any moment to swarm in with demands for shedloads of arbitrary-seeming data. Just as a news scandal can conjure a flashmob of millions of net-users to Google's homepage, all searching for 'tsunami' or 'paris hilton', an exciting discovery in genetics sends the whole bioinformatics community to the Sanger's servers, all demanding the same thing.

You can't go far in this world without some sort of comparison to Google, which is the biggest of the big gorillas. How big, though, is not quite clear — and nor is it clear how it manages its flashmobs and other petaproblems. In the ten years since the company's founding, Google's data-serving systems have gone from a set of commodity PCs connected to a hard-disk array built into an enclosure made from Lego bricks to a global system of data farms unmatched by anyone else. Each of those data centres is designed from the foundation up to operate as a single big computer. Google buys components optimized for the kind of operations that relentlessly hammer its servers. It has software for the job such as Google File System (which distributes three copies of each piece of information in a way that makes it easy to recover when the inevitable failure occurs) and Google MapReduce, a system for automatically and efficiently making large data sets amenable to parallel processing. Google's distinguished engineers present papers at learned conferences explaining in detail how it all works. People such as Butcher pay close attention.

But then there's the closed part: all the specific metrics and data-porn that Google considers of competitive significance. The company no longer says how big the model, or copy, of the web's material spread through its data centres is. It doesn't disclose the dimensions or capacity of those data centres. *Nature* wanted me to visit one for this piece, but a highly placed Googler told me that no one from the press had ever been admitted to a Google data centre; it would require a decision taken at the board level. Which is too bad. But it's not as if the world is bereft of other computer installations with mind-bending requirements. And at CERN, the Sanger and XS4ALL in the Netherlands, I found myself welcomed into the roaring sanctums of computing, escorted around by sysadmins eager to show off the hellaciously complex, monstrously powerful machines that they've been able to put together and put to use.

Repository of all knowledge

The primary XS4ALL facility at the World Trade Centre near Schiphol Airport in Amsterdam was actually built to house a mighty array of huge telephone switches in 2000. KPN, the then Dutch national telecom company, fitted it out generously, with several months' worth of diesel in subterranean tanks for its uninterruptible power supply's back-up generators, and two independent cooling systems, with one raised two storeys off the ground to flood-proof it (this is the lowlands after all). But telecom deregulation was not kind to KPN, and the switches never came. So now the facility houses XS4ALL, a once-notorious Dutch Internet service provider that has somehow made it bigtime. Hacktic, a collective of hackers, established XS4ALL in 1993 to help cover the costs of the Internet link they had set up. In 1994, KPN shut down all of XS4ALL's lines after the collective published

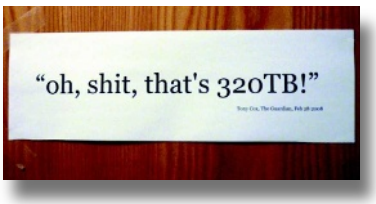
C. DOCTOROW

The XS4ALL building in Amsterdam (top left); the back of stacks at XS4ALL (bottom left) and CERN near Geneva (facing page, top); a PetaBox (facing page, bottom); and Tony Cox's door (inset).

C. DOCTOROW

an article explaining how to cheat the phone company's punitive long-distance tolls — the ISP came back online only after posting a 60,000 guilder (US$35,000) cash bond. Just four years later, after XS4ALL had grown into one of the most successful ISPs in the lowlands, the state company bought out its former gadfly. Today XS4ALL is as independent as a subsidiary of a former government monopoly can be, but its members are not above sharing digs with their corporate parent, especially as the corporate parent is such a spendy sort of sugar-daddy. XS4ALL has taken over two storeys of the would-be switching centre with hackerish humour: the raised floors sport Perspex panels revealing neon lights and jumbles of entombed PC junk; there is a chill-out room for sysadmins who come on-site to run backups or swap drives; a poster listing the facility's regulations ends: "Rule 12: No sex in the data centre."

C. DOCTOROW



The mix of freewheeling hacker humour, deadly serious commitment to free speech and solid technological infrastructure made XS4ALL a natural choice to host a mirror copy of the Internet Archive (archive.org), a massive 'repository of all knowledge'. The archive's best-known feature is the Wayback Machine, an archive of most of the public pages on the World Wide Web that allows visitors to 'travel in time' and see what any given URL looked like on any given date since 1996. But it also serves as a repository for practically every public domain and Creative Commons-licensed digital document it can lay its hands on. It is the brainchild of philanthropist Brewster Kahle — co-creator of Wide Area Information Servers, or WAIS, one of the first Internet search engines — who wants it to provide "universal access to all knowledge". In a world of here today/gone tomorrow Web 2.0 companies willing to host your video, pictures or text, the archive stands out as a sober-sided grown-up with a commitment to long-term (infinite-term) hosting.

Inside the XS4ALL data centre, which is about the size of a football pitch, my hosts took me past aisle after aisle of roaring machines to the Internet Archive's European mirror.

"That's it, huh?"

Two racks, each the size of a modest refrigerator, each holding north of a petabyte's worth of information. These are the PetaBoxes, the Internet Archive's web-in-a-box systems. Designed as a shippable backup for every freely shareable file anyone cares to upload and hundreds of copies of everything else, too, they betray the archive's US origins in the strip of American-style electric outlets running down one strut, a column of surprised clown-faces Fed-Exed from across the ocean. A couple of other things set them apart. Each rack draws 12 kilowatts, whereas a normal rack at the facility draws 4.5 kilowatts; the drive-housings are covered in a rather handsome fire-engine-red enamel. Apart from that, the PetaBoxes are just another pair of racks.

Yet housed in these machines are hundreds of copies of the web — every splenetic message-board thrash; every dry e-government document; every scientific paper; every pornographic ramble; every libel; every copyright infringement; every chunk of source code (for sufficiently large values of 'every', of course).

They have the elegant, explosive compactness of plutonium.

Something dawns on me: I ask my XS4ALL tour guides, shouting over the jet-engine roar: "If there are all those copies of the web on the PetaBoxes, what's in all the other machines?"

"Oh, customer stuff. Intranets. Databases. E-mail. Usenet." In other words, all the dynamic stuff, the private stuff, the dark web that is invisible to search engines, and all the processor power needed to serve it. All the reasons that Google can't exist just in a couple of red PetaBoxes.

"How does KPN feel about housing these two extraordinary boxes?"

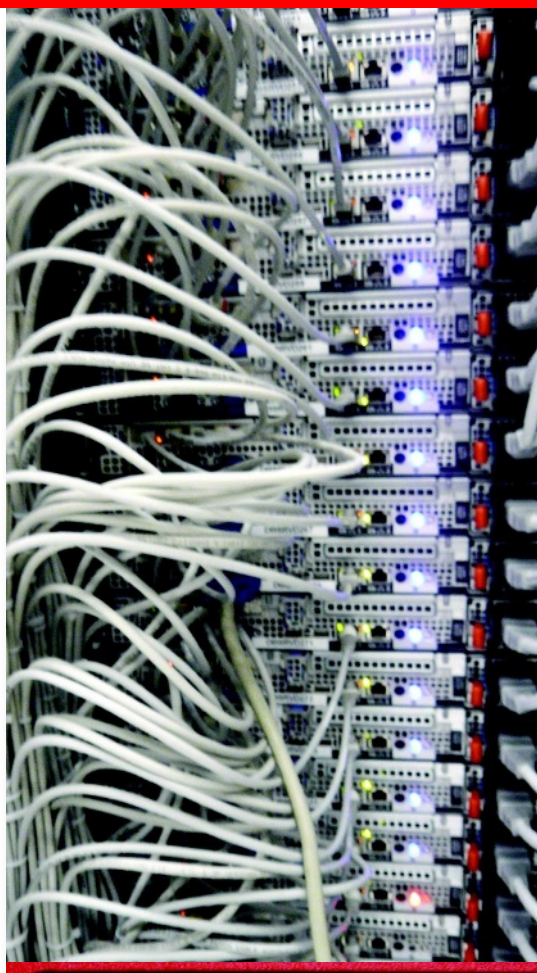
"Oh," they say, exchanging a mischievous glance, "I don't think they know we have them here."

In a data centre such as this, a working approximation of 'all knowledge' can be slipped into the cracks like a 32-MB memory card jingling in my pocket.

600 million collisions a second

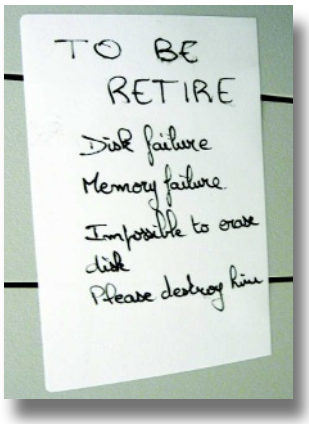
The archive has three real-time mirrors: the original in San Francisco's Presidio, just south of the Golden Gate, the XS4ALL mirror, and a third under the New Library of Alexandria in Egypt. A keen observer will note that these are variously placed on the San Andreas Fault, in a flood-zone, and in a country with a 27-years-and-running official 'state of emergency' that gives the government the power to arbitrarily restrict speech and publication. Someone needs to buy Kahle a giant cave in Switzerland. Like the one I'm off to now, which will be housing the data from the biggest experiments on the most powerful machine ever conceived.

Except it turns out that the data centre at CERN is less hall of the mountain king and more high-school gymnasium. The caverns measureless to man through which the



LHC runs are reserved for making the data. The systems storing them have much more humdrum quarters. The slight sense of anticlimax is emphasized by the unflappable calm of Tony Cass, CERN's leader of fabric infrastructure and operations; his data centre may be about to become the white-hot focus of the entire world's high-energy physics community, but Cass is surprisingly and perhaps bit disappointingly relaxed. Indeed, when we met just a few weeks before the LHC was about to see its first circulating beam, on 10 September, he was headed off on holiday.

Built in the 1970s to be the only data centre that CERN would ever need, Cass's current facility is now just a stopgap on the way to the construction of a bigger, faster centre that will absorb 15 petabytes a year of experimental data from the LHC. Although the rack after rack of systems in the current centre are nearly new, there are already plans to replace them. The basement is a graveyard of already-replaced generic PCs that are slowly being cleansed of data and shipped to bidders from the former Soviet Union.



C. DOCTOROW

The difference between Cass's challenges and Butcher's is a difference in the way that physics and biology work. At the Sanger, the charge-coupled devices (CCDs) in the sequencers can vomit out TIFF image files by the terabyte around the clock, but they are useless until processed, analysed and shrunk down to a far more manageable summary of what those vast image files actually meant. The original data are thrown away — the Sanger is confident that there will never be anything new to be learned from looking at the raw image files later. And there would be no way of keeping it except on tape, and tape, as Butcher will tell you, is slow, impractical and failure prone. As a former sysadmin myself, I can attest to the inherent tetchiness of tape. The Sanger reduces the images to more amenable data and then sends everything off to various mirror sites using a custom-made file-transfer protocol implemented over what's known as user datagram protocol (UDP); this allows the gene genies to saturate entire transoceanic links without having to wait for any of the finicky TCP handshaking and error-correction nonsense used for Internet traffic. It's slow compared with CERN's approach — CERN leases its own fibre, at great expense — but it certainly beats the old-school open-source system used to glue together the Internet Archive's mirrors.

C. DOCTOROW

If only high-energy physics were so amenable to throwing stuff away, Cass's life would be a lot simpler. It's not. The meaning of a sequencer run is pretty straightforward, and won't change. The meaning of a particle collision is continuously reassessed based on new information about the instrument's performance. Physicists will want to reanalyse all the collisions expected from the LHC from the first to the last. Six hundred million events a second for year after year, analysed over and over again as the physicists' models become more refined. And that means a lot of storage — the kind of storage you can't load onto a reasonable quantity of spinning drives. The kind of storage you need to put on tape.

I am, admittedly, prone to swooning over a well-designed bit of IT kit, but I have never developed as deep and meaningful and instantaneous a relationship as the one I formed with the two tape-loading robots in the basement of the CERN data centres.

The Vader-black machines, one built by StorageTek, a subsidiary of Sun Microsystems, the other by IBM, are housed in square, meshed-in casings the size of small shipping containers. From within them comes a continuous clacking noise like the rattling of steel polyhedral dice on a giant's Dungeons & Dragons table. I pressed my face against the mesh and peered in fascination at the robot arms zipping back and forth with tiny, precise movements, loading and unloading 500-GB tapes with the serene grace of Shaolin monks. Did I say tape is tetchy? I take it back. Tape is beautiful.

Each robot-librarian tends 5 PB of data. It will jump shortly to 10 PB each when the 500-GB tapes are switched to 1-TB models — an upgrade that will take a year of continuous load/read/load/write/discard operations, running in the interstices between the data centre's higher-priority tasks. When that is done, there should be 2-TB tapes to migrate to, bringing the two robots' total up to 40 PB. At least, that's what CERN hopes.

The tape libraries will allow the regular reassessment of the LHC data — unloading, reprocessing and reloading all the data on each of the tapes. A complete reprocessing will take a year, in part because, although it is higher priority than migrating the data to higher-density tapes, it still takes a backseat to the actual science —

C. DOCTOROW

A tape robot at CERN (top left); cooling devices at XS4ALL (bottom left); a label (above) for obsolete CERN computers; and rescue disks at CERN (facing page).

to jobs requested from anywhere in the world.

CERN embodies borderlessness. The Swiss–French border is a drainage ditch running to one side of the cafeteria; it was shifted a few metres to allow that excellent establishment to trade the finicky French health codes for the more laissez-fair Swiss jurisdiction. And in the data sphere it is utterly global.

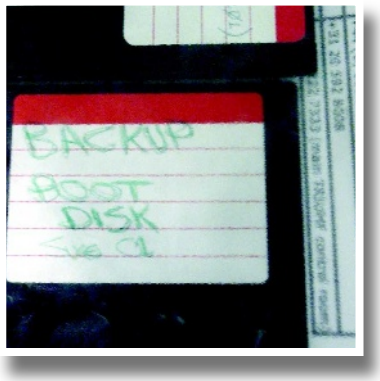
Cass's operation is backstopped by ten 'Tier One' facilities around the world that replicate its tape library, and some hundreds of 'Tier Two' facilities that provide compute-power to operate on those data, all linked by dedicated, high-speed fibre, part of a global network that attempts to tie the world's high-energy physics institutions into a single, borderless facility. A researcher who logs into the CERN data centre runs code without worrying which processors execute it or which copy of the data it is run on. The birthplace of the web, which demolished borders for civilians, CERN is ushering in a borderless era for data-intensive science, an era in which US researchers run code on Iranian supercomputers and vice versa, without regard for their respective governments' sabre rattling. Cass wants to weld the world's physics computers into a single machine.

Sysadmin nightmares

At each data centre I asked the sysadmins for their worst fears. Universally, the answer was heat. Data centres are laid out in alternating cool and hot aisles, the cool looking at the front of the racks, the hot at the back. At CERN, they actually glass over the cool aisles to lower the cooling requirements, turning them into thrumming walk-in fridges lined with millions of tiny, twinkling lights.

If power is cut to the cooling system in one of these places, you've got minutes for a clean shutdown of the systems before their heat goes critical. XS4ALL has a particularly impressive cooling system, a loop that runs from the 5 °C, 30-metre depths of nearby Lake Nieuwe Meer, warms to 16 °C in the centre's exchangers, and then plunges back to the lake-bottom to be cooled again. The site manager Aryan Piets estimates that if it broke down and the emergency system didn't come on, the temperature in the centre would hit 42 °C in ten minutes. No one could cleanly bring down all those machines in that time, and the dirtier the shutdown, the longer the subsequent start-up, with its rebuilding of databases and replacement of crashed components. Blow the shutdown and stuff starts to melt — or burn.

Data centres do face more exotic risks. Google once lost its transoceanic connectivity because of shark bites. Butcher lives in fear of a Second World War fighter plane going astray from the airshows at nearby Duxford airfield and crashing into the Ice Box. At CERN they worry about people believing the worries that the Universe will wink out of existence when they fire up the LHC. But the real worry is power and its management. Data centres built in the giddy dotcom heyday assumed that racks would sport one processor core per unit and planned cooling and energy accordingly. But that is not the way the technology has gone. Computers have got faster not through faster cores, but through more of them. With 16 cores or more per unit, data centres around the world sit half-empty, unable to manage the power-appetites of a whole room's worth of 2008's ultra-dense computing. And everyone lives in fear of the electrical fault that sparks a blaze that wafts killing soot into the hungry ventilation intakes on the racks.



C. DOCTOROW

A big part of the problem — and possibly of its solution — is that most of a data centre's compute capacity is idle much of the time. No sysadmin would provision a data centre to run at capacity around the clock, lest it melt down (along with the sysadmin's career) the first time something really juicy increased the load. Yet whether a network card is saturated or idle, it still burns 100% of its energy draw. The same with video cards, power supplies, RAM and every other component except for some CPUs. So these idle systems whir away, turning coal into electricity into heat that has to be cooled with coal turned into electricity turned into heat, and the planet warms and the bills soar. Every decibel of noise roaring through the centres is waste, energy pissed away for no benefit.

> **"If the emergency system didn't come on the temperature would hit 42 °C in ten minutes."**

The people with the biggest data centres have the biggest problem — and the biggest resource to throw at it. Google buys its systems in enough bulk that it can lay down the law to component suppliers, demanding parts that draw power proportional to the amount of work that they are doing. Its holistic approach to the data centres, treating each one as a single PC, means that it can plan for idleness and peak load alike, and keep the energy bills under control. Everyone agrees that something like this is the way forward, that the future of data centres must be cooler, and quieter.

That said, a certain discomforting noise has its advantages. "I don't want it to ever get too comfortable in here," says Cass. "I like it that people access us remotely. It just doesn't scale, having every scientist drop in to run jobs here in the centre."

And if Google leads the way because it has to feed people's need for Paris Hilton searches and peeks at their own roof on Google Earth, that is quite fitting. Whereas scientists unzip new genomes and summon new particles from the roiling vacuum with technologies beyond compare, the secret of data storage and processing is a lot simpler: commodity components. There is a huge ingenuity in how you use them, cool them, arrange them and keep them from melting, but the basic ingredients of a petacentre are the ingredients of life on the net. Everything I've seen on these trips was basically made out of the same stuff I've got lying around the flat. Gene-sequencers use multi-megapixel CCDs — cheap and cheerful in this era of digital photography — to generate TIFFs that I could open with the open-source image-manipulation program that came with my free Ubuntu GNU/Linux operating system. The hard-drives in the server cases are the same cheap, high-capacity Seagates and Toshibas that I have in the little box I stuck under the stairs and wired up to my telly to store away a couple of terabytes of video, audio and games.

A decade ago, a firm's 'mainframe' was a powerful beast made from specialized components never seen outside the cool, dust-free environs of the data centre. Today, mainframe is more synonymous with the creaky old legacy system that no one can be bothered to shut down because it is running an obscure piece of accounting software that would be a pain to port to a modern system. The need for special hardware just isn't there any more. Even Google's 'energy-proportional' future is just an expansion of the power-management and heat-dissipation technology developed for laptops, and any gains achieved on the server side will also come to our desktops. I've got everything I need lying around the office to make my own petacentre — I just need more of it. And a much bigger fridge. Or a cool-bottomed lake.

That said, I don't have a tape robot.

But I really, really want one. ■

Cory Doctorow is a digital-rights activist, author and co-editor of Boing Boing, a blog. His most recent novel is *Little Brother*.
See Editorial, page 1.