Blog Home

Feed RSS

The developers of Backblaze post on this weblog about online backup, keeping data safe and other rants about life. Visit the How it Works page to learn more.

Tags

Backblaze for Business

Backblaze Fun

Backblaze Tips & Tricks

Backup Awareness Month

Backup Devices

Backup Needs

Backup News

Cloud Storage

Customer Stories

Data Loss

Events

Jobs

Kudos

Locate My Computer

Petabytes on a Budget v2.0: Revealing More Secrets

Tim Nufire July 20, 2011

It's been over a year since Backblaze revealed the designs of our first generation (67 terabyte) storage pod. During that time, we've remained focused on our mission to provide an unlimited online backup service for $5 per month. To maintain profitability, we continue to avoid overpriced commercial solutions, and we now build the Backblaze Storage Pod 2.0: a 135-terabyte, 4U server for $7,384. It's double the storage and twice the performance—at lower cost than the original.

In this post, we'll share how to make a 2.0 storage pod, and you're welcome to use the design. We'll also share some of our secrets from the last three years of deploying more than 16 petabytes worth of Backblaze storage pods. As before, our hope is that others can benefit from this information and help us refine the pods. (Some of the enhancements are contributions from helpful kindred pod builders, so if you do improve your Backblaze pod farm, please balance the Karma and send us your suggestions!)

Quick Review - What makes a Backblaze Storage Pod

A Backblaze Storage Pod is a self-contained unit that puts storage online. It's made up of a custom metal case with commodity hardware inside. You can find a parts list in Appendix A. You can also link to a power wiring diagram, see an exploded diagram of parts, and check out a half-assembled pod. The two most noteworthy factors are that the cost of the hard drives dominates the price of the overall pod and that the system is made entirely of commodity parts. For more background, read the original blog post. Now let's talk about the changes.

Density Matters - Double the Storage in the Same Enclosure

We upgraded the hard drives inside the 4U sheet metal pod enclosure to store twice as much data in the same space. After the cost of filling a rack with pods, one datacenter

Mac Love
Offers
Release
Startup Life
Storage Pod
Tech News
TechBytes
Uncategorized

rack containing 10 pods costs Backblaze about $2,100 per month to operate, roughly divided equally into thirds for physical space rental, bandwidth, and electricity. Doubling the density saves us half of the money spent on both physical space and electricity. The picture below is from our datacenter, showing 15 petabytes racked in a single row of cabinets. The newest cabinets squeeze one petabyte into three-quarters of a single cabinet for $56,696.

Our online backup cloud storage is our largest cost, and we are obsessed with providing a service that remains secure, reliable and, above all, inexpensive. We've seen competitors unable to react to these demands who were forced to exit the market, like Iron Mountain, or raise prices, like Mozy and Carbonite. Controlling the hardware design has allowed us to keep prices low.

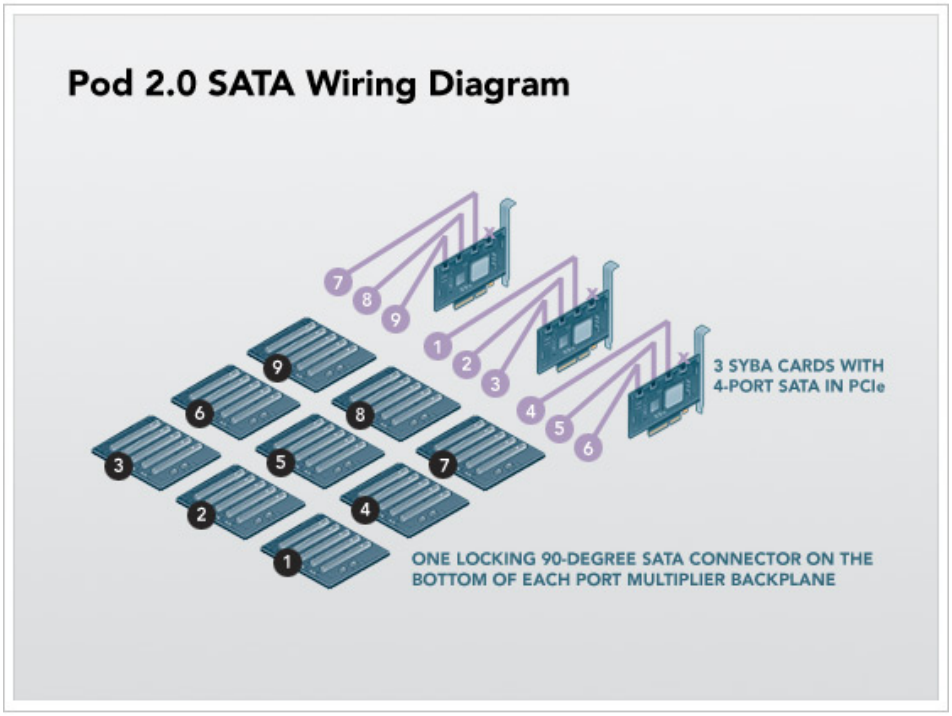
We are constantly looking at new hard drives, evaluating them for reliability and power consumption. The Hitachi 3TB drive (Hitachi Deskstar 5K3000 HDS5C3030ALA630) is our current favorite for both its low power demand and astounding reliability. The Western Digital and Seagate equivalents we tested saw much higher rates of popping out of RAID arrays and drive failure. Even the Western Digital Enterprise Hard Drives had the same high failure rates. The Hitachi drives, on the other hand, perform wonderfully.

Twice as Fast

We've made several improvements to the design that have doubled the performance of the storage pod. Most of the improvements were straightforward and helped by Moore's Law. We bumped the CPU up from the Intel dual core CPU to the Intel i3 540 and upgraded the motherboard from one Gigabit Ethernet port to a Supermicro motherboard with two Gigabit Ethernet ports. RAM dropped in price, so we doubled it to 8 GB in the new pod. More RAM enables our custom Backblaze software layer to create larger disk caches that can really speed up certain types of disk I/O.

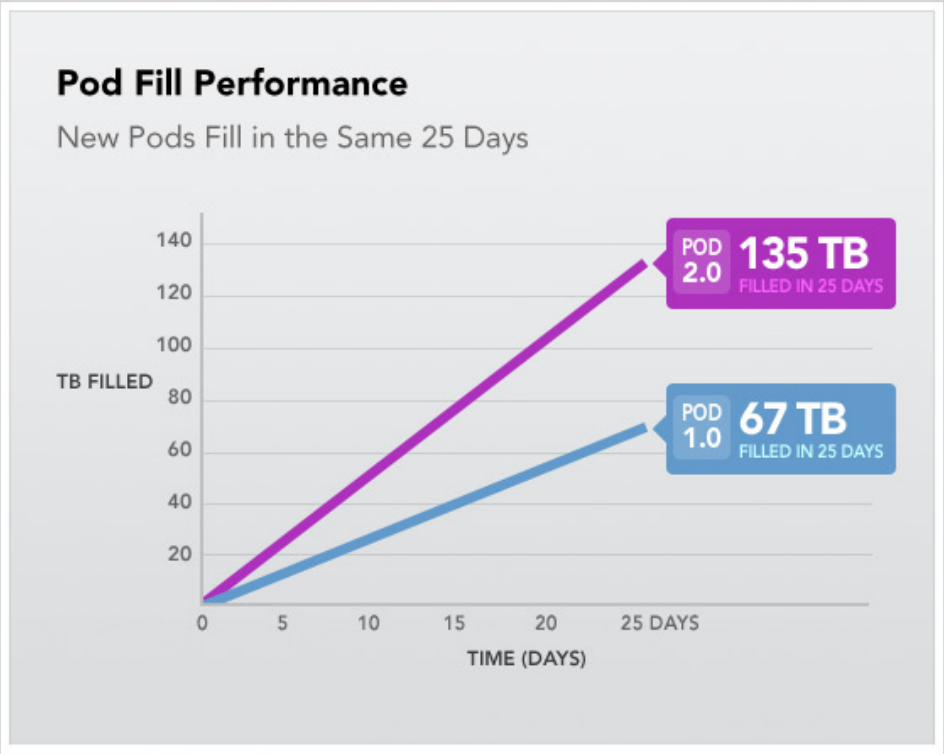
In the first generation storage pod, we ran out of the faster PCIe slots and had to use one slower PCI slot, creating a bottleneck. Justin Stottlemyer from Shutterfly found a better PCIe SATA card, which enabled us to reduce the SATA cards from four to three. Our upgraded motherboard has three PCIe slots, completely eliminating the slower PCI bottleneck from the system. The updated SATA wiring diagram is seen below. Hint: The

pod will work if you connect every port multiplier backplane to a random SATA connection, but if you wire it up as shown below, the 45 drives will appear named in sequential order.



We upgraded the Linux 64-bit OS from Debian 4 to Debian 5, but we no longer use JFS as the file system. We selected JFS years ago for its ability to accommodate large volumes and low CPU usage, and it worked well. However, ext4 has since matured in both reliability and performance, and we realized that with a little additional effort we could get all the benefits and live within the unfortunate 16 terabyte volume limitation of ext4. One of the required changes to work around ext4's constraints was to add LVM (Logical Volume Manager) above the RAID 6 but below the file system. In our particular application (which features more writes than reads), ext4's performance was a clear winner over ext3, JFS, and XFS.

With these performance improvements, we see the new storage pods in our datacenter accepting customer data more than twice as fast as the older generation pods. It takes approximately 25 days to fill a new pod with 135 terabytes of data. The chart below shows the measured fill rates of an old Pod versus a new Pod, both under real-world maximum load in our datacenter.



Please note: The above graph is not the benchmarked write performance of a pod; we have easily saturated the Gigabit pipes copying data from one pod to another internally. This graph shows pods running in production, accepting data from thousands of simultaneous and independent desktop machines running Windows and Mac OS, where each desktop is forming HTTPS connections to the Tomcat web server and pushing data to the pod. At the same time, as customers are preparing restores that read data off those drives, there are system cleanup processes running, occasional RAID repairs, etc. In this end-to-end measurement, the new pods are twice as fast in our environment.

Lessons Learned: Three Years, 16 Petabytes and Counting

Backblaze is employee owned (with no VC funding or other deep pockets), so we have two choices: 1) stay profitable by keeping costs low or 2) go out of business. Staying profitable is not just about upfront hardware costs; there are ongoing expenses to consider.

One of the hidden costs to a datacenter is the headcount (salary) for the employees who deploy pods, maintain them, replace bad drives with good, and generally manage the facility. Backblaze has 16 petabytes and growing, and we employ one guy (Sean) whose fulltime job is to maintain our fleet of 201 pods, which hold 9,045 drives. Typically, once every two weeks, Sean deploys six pods during an eight-hour work day. (He gets a little help from one of us to lift each pod into place because they each weigh 143 pounds.)

Our philosophy is to plan for equipment failure and build a system that operates in spite of it. We have a lot of redundancy, ensuring that if a drive fails, immediate replacement isn't critical. So at his leisure, Sean also spends one day each week replacing drives that have gone bad. As of this week, Backblaze has more than 9,000 hard drives spinning in the datacenter, the oldest of which we purchased four years ago. We see fairly high infant mortality on the hard drives deployed in brand new pods, so we like to burn the pods in for a few days before storing any customer data. We have yet to see any drives

die because of old age, which will be fascinating to monitor in the next few years. All told, Sean replaces approximately 10 drives per week, indicating a 5 percent per year drive failure rate across the entire fleet, which includes infant mortality and also the higher failure rates of previous drives. (We are currently seeing failures in less than 1 percent of the Hitachi Deskstar 5K3000 HDS5C3030ALA630 drives that we're installing in pod 2.0.)

We monitor the temperature of every drive in our datacenter through the standard SMART interface, and we've observed in the past three years that: 1) hard drives in pods in the top of racks run three degrees warmer on average than pods in the lower shelves; 2) drives in the center of the pod run five degrees warmer than those on the perimeter; 3) pods do not need all six fans—the drives maintain the recommended operating temperature with as few as two fans; and 4) heat doesn't correlate with drive failure (at least in the ranges seen in storage pods).

One important note: Because all of the parts (including drives) in the Backblaze storage pod come with a three-year warranty, we rarely pay for a replacement part. The drive manufacturers take back failed drives with "no questions asked" and send free replacements. If you figure that storage resellers, such as NetApp and EMC, tack on a three-year support fee, a petabyte of Backblaze storage costs less than their support contract alone. A chart below takes all of our experience into account and shows what it costs to own and maintain a Petabyte of storage for three years:

In the chart above, the economies of scale only kick in if you really do need to store a full petabyte or more. For a small amount of data (a few terabytes), Amazon S3 could easily save money, but the Amazon option is clearly a dubious financial choice for a company with large, multi-petabyte storage needs.

Final Thoughts

The Backblaze storage pod is just one building block in making a cloud storage service. If all you need is cheap storage, this may suffice. If you need to build a reliable, redundant, monitored storage system, you've got more work ahead of you. At Backblaze we've developed software that manages and monitors the cloud service, proprietary technology that we've developed over the years.

We offer our storage pod design free of any licensing or any future claims of ownership. Anybody is allowed to use and improve upon it. You may build your own cloud system and use the Backblaze storage pod as part of your solution. The steps to assemble a storage pod, including diagrams, can be found on our original blog post, and an updated list of parts is provided below in Appendix A. We don't sell the design, so we don't provide support or a warranty for people who build their own. To all of those builders who take up the challenge, we'd love to hear from you and welcome any insights you provide about the experience. And please send us a photo of your new 135 Terabyte pod.

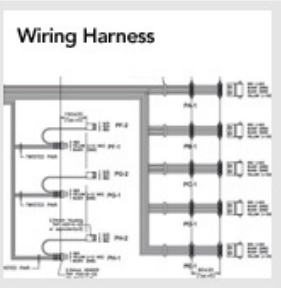
Appendix A – Price List:

Item	Qty	Price	Total
3 Terabyte Drives Hitachi 3TB 5400 RPM HDS5C3030ALA630	45	$120.00	$5,400
4U Custom Case (Available in quantities of 1 from Protocase for $875) – link to 3D design	1	$350	$350
760 Watt Power Supply Zippy PSM-5760 Power Supply	2	$270	$540
Port Multiplier Backplanes Available in qty of 9 for $47 from (CFI Group) CFI-B53PM 5 Port Backplane (Sil3726)	9	$41	$369
Intel i3 540 3.06 Ghz CPU	1	$110	$110
Port PCIe SATA II Card Syba PCI Express SATA II 4 x Ports RAID Controller Card SY-PEX40008	3	$50	$150
Motherboard SuperMicro MBD-X8SIL-F-B	1	$154	$154
Case Fan Mechatronics G1238M(OR E)12B1-FSR 12V 3-Wire Fan	6	$12	$70
8GB DDR3 RAM Crucial CT25672BA1339 2GB, DDR3 PC3-10600 (4x 2GB = 8GB total)	2	$58	$116
160 GB Boot Drive Western Digital Caviar Blue WD1600AAJS 160GB 7200 RPM	1	$39	$39
On/Off Switch FrozenCPU ele-302 Bulgin Vandal Momentary LED Power Switch 12" 2-pin	1	$30	$30
SATA II Cable Newegg GC36AKM12 3 Foot SATA Cable	9	$2	$18
Nylon Backplane Standoffs Fastener SuperStore 1/4" Round Nylon Standoffs Female/Female 4-40 x 3/4"	72	$.18	$13
HD Anti-Vibration Sleeves Aero Rubber Co. 3.0 x .500 inch EPDM (0.03" Wall)	45	$.23	$10
Power Supply Vibration Dampener Vantec VDK-PSU Power Supply Vibration Dampener	2	$4.5	$9

Fan Mount (front) Acoustic Ultra Soft Anti-Vibration Fan Mount AFM02	12	$.18	$2
Fan Mount (middle) Acoustic Ultra Soft Anti-Vibration Fan Mount AFM03	12	$.18	$2
Nylon Screws Small Parts MPN-0440-06P-C Nylon Pan Head Phillips 4-40 x 3/8″	72	$.02	$1
Foam Rubber Pad House of Foam 16″ x 17″ x 1/8″ Foam Rubber Pad	1	$1	$1

TOTAL: $7,384

Custom wiring harnesses for PSU1 and PSU2 (the Zippy power supplies):
See detailed wiring harness diagrams.



SATA Chipsets
SiI3726 on each port multiplier backplane to attach five drives to one SATA port.
SiI3124 on three PCIe SATA cards. Each PCIe card has four SATA ports on it, although we only use three of the four ports.

Tags: Cloud Storage, Storage Pod

W Kent Kovac · Michigan State University
Just finished building two of your older version at the Plant Research Laboratory at Michigan State, using them both as frontends into a ROCKS cluster, great stuff!
3 · Like · Reply · Moderate · Subscribe · Wednesday at 2:17pm

View 1 more

Chris Gulvik · Wisc Oshkosh
Sweet! Now all that is left is to write a script to generate well-written proposals for NSF funding submissions by using adaptive learning from reviewing the literature on a topic query to address new and 'pressing' biological questions. Once complete, you could put it into the cluster and soon have a lab/'army' larger than Venter's with all of the funding :) Oh, and make it hypothesis-based!
2 · Like · Reply · Moderate · Wednesday at 8:48pm

Sean O'Malley · East Lansing, Michigan
It is sweet! If you tack on a Marvell Dragonfly to your server hosts, it would probably rock for a VM cluster too.
Like · Reply · Moderate · 6 hours ago

W Kent Kovac · Michigan State University

Hmm perhaps... are you a MSUer?

Like · Reply · Moderate · 4 hours ago

Brian Graves

why ext4 and not zfs?

3 · Like · Reply · Moderate · Subscribe · Wednesday at 9:22am

View 7 more

Elliott Sims · Site Reliability Engineer at Facebook

There's a difference between using a potentially-unstable communications library that can be replaced/reverted without impact and using a potentially-unstable FS that in the event of a problem eats your customers' data.

1 · Like · Reply · Moderate · Yesterday at 1:40pm

Gleb Budman · Top Commenter · CEO at Backblaze

Logan - perfectly reasonable, but there isn't a strong incentive from us to switch from Linux. For someone building a new system, might work great.

Adam - since we subdivide pods into volumes anyways, the 16TB limit is not a huge deal, but good to know you've liked Btrfs.

Like · Reply · Yesterday at 2:01pm

Brent Garber · Arkansas

Not denying there isn't a difference, just saying that version numbers have absolute zero relevance to quality and stability, so going 'it's not even 1.0' is a silly argument to make.

1 · Like · Reply · Moderate · 22 hours ago

Alan Aspuru-Guzik · Harvard

We are going to build our second Backblaze here at Harvard for the http://cleanenergy.harvard.edu Clean Energy Project. We will check our notes with this one! We did a 90TB variation recently.

2 · Like · Reply · Moderate · Subscribe · Wednesday at 8:46am

Gleb Budman · Top Commenter · CEO at Backblaze

We'll look forward to hearing how it goes!

Like · Reply · Wednesday at 11:59am

Paul D. Walker

What are the chances of selling the cases and parts without hard drives to potential customers?

2 · Like · Reply · Moderate · Subscribe · Wednesday at 8:44pm

Larry Wright · South Grand Prairie High School

Protocase already does this for $5k.

Like · Reply · Moderate · Yesterday at 12:26pm

Simon White · Newtownards

Would love one of those for no reason :D

1 · Like · Reply · Moderate · Subscribe · Wednesday at 8:47am

Gleb Budman · Top Commenter · CEO at Backblaze

Kind of link Don Honabach, who built one of the v1.0 type to store all his movies? http://blog.backblaze.com/2009/10/12/user-builds-extreme-media-server-based-on-a-backblaze-storage-pod/

3 · Like · Reply · Wednesday at 12:02pm

Simon White · Newtownards

I have a media server... 3TB in it. This puts it to shame.

Like · Reply · Moderate · Wednesday at 12:07pm

Bruce Hazan · EPITECH

Great work guys!

1 · Like · Reply · Moderate · Subscribe · Wednesday at 11:46am

Gleb Budman · Top Commenter · CEO at Backblaze

Thank you!

Like · Reply · Wednesday at 12:13pm

Andrew DeSio · SPaRtAn at Sparta

Just curious, but is cost the only reason your boot drive isn't an SSD? I suppose the reliability would be nice, but you could replace a HDD several times for the price of the SSD. Curious to hear your take on it.

Like · Reply · Moderate · Subscribe · Wednesday at 3:21pm

Gleb Budman · Top Commenter · CEO at Backblaze

Absolutely. We have considered using an SD card, which would be a wash in terms of price, but seemed more complicated. Even on reliability, the jury seems to still be out.

Like · Reply · Wednesday at 4:50pm

Peter Kimball · Hamilton College

Can you give us an idea of the max and typical power draws per Pod? We're definitely interested in building a few, just trying to get a handle on what sort of power upgrade we'd need in our cabinet...

Like · Reply · Moderate · Subscribe · Wednesday at 12:43pm

Gleb Budman · Top Commenter · CEO at Backblaze

About 500 watts for the low-power drives, 625 for high-power. If you build some pods, would love to know how you use them!

Like · Reply · Wednesday at 12:49pm

Peter Kimball · Hamilton College

We'll definitely share the results, thanks for the info!

Like · Reply · Moderate · Wednesday at 12:59pm

Willie Slepecki

so you guys are using a software controlled RAID 6 instead of a hardware based controller it looks like. what programs are you using to create the RAID array?

Like · Reply · Moderate · Subscribe · Wednesday at 4:03pm

Gleb Budman · Top Commenter · CEO at Backblaze

We're using mdadm (multi-disk admin) - the Linux software RAID package.

Like · Reply · Wednesday at 4:51pm

Matt Keenan · University of Technology, Sydney

Just curious why you don't use mdadm raid10 with far layout 3? Wouldn't this still give you +2 reliability without the CPU overhead?

Like · Reply · Moderate · Yesterday at 8:57am

Willie Slepecki

second question, what are you using to monitor the health of each drive, meaning how do you know when a drive fails.

Like · Reply · Moderate · Subscribe · Wednesday at 4:05pm

Gleb Budman · Top Commenter · CEO at Backblaze

To monitor the drives we use the Debian package "smartmontools", mdadm to monitor the RAID arrays, and Zabbix for alerts/trending. Above that, we have developed an entire web-based admin system to manage our entire cloud storage farm.

Like · Reply · Wednesday at 4:58pm

Justin Mecham · Veazie, Maine

Zabbix is win; speaking from experience. It takes a little while to get it dialed in and alerting on what you want, but once you do it's excellent.

Like · Reply · Moderate · Yesterday at 3:09pm

View 41 more

« Backblaze fully supports OS X Lion (10.7) -