

GreenFS

Making Enterprise Computers Greener by Protecting Them Better

Nikolai Joukov¹ Josef Sipek²

¹IBM T.J. Watson Research Center
Hawthorne, NY, USA

²Computer Science Department
Stony Brook University

EuroSys European Conference on Computer Systems, 2008

Outline

- 1 Motivation
 - Enterprise Concerns
 - Existing Solutions
- 2 GreenFS
 - Goals
 - Design
 - Implementation
- 3 Evaluation

Power Consumption

- Office systems constitute 1% of total US power
- Data centers consume an additional 1.2%
- Power constitutes ~50% of the TCO of a data center
- Hard drives consume 5-31% of total system power

Power Consumption

- Office systems constitute 1% of total US power
- Data centers consume an additional 1.2%
- Power constitutes ~50% of the TCO of a data center
- Hard drives consume 5-31% of total system power

Power Consumption

- Office systems constitute 1% of total US power
- Data centers consume an additional 1.2%
- Power constitutes ~50% of the TCO of a data center
- Hard drives consume 5-31% of total system power



Power Consumption

- Office systems constitute 1% of total US power
- Data centers consume an additional 1.2%
- Power constitutes ~50% of the TCO of a data center
- **Hard drives consume 5-31% of total system power**

Power Consumption

- Office systems constitute 1% of total US power
- Data centers consume an additional 1.2%
- Power constitutes ~50% of the TCO of a data center
- Hard drives consume 5-31% of total system power

	Unit cost (\$)	Power (W)	Power cost (\$/year)*	5 year cost (\$)
Desktop HDD	\$100	10	\$18	\$90
Server HDD	\$250	30	\$53 [†]	\$265

* Assumes \$0.20/kWh

[†] Includes cost of cooling and UPS losses

Additional Concerns

Noise

- Office noise reduces productivity, motivation, ability to learn
- Desktop hard drives produce levels between 30 - 50 dBA

Additional Concerns

Noise

- Office noise reduces productivity, motivation, ability to learn
- Desktop hard drives produce levels between 30 - 50 dBA

Additional Concerns

Noise

- Office noise reduces productivity, motivation, ability to learn
- Desktop hard drives produce levels between 30 - 50 dBA

Shocks

- Spinning hard drives are fragile and sensitive to shocks
- Disk head and platter collisions result in rapid data loss

Additional Concerns

Noise

- Office noise reduces productivity, motivation, ability to learn
- Desktop hard drives produce levels between 30 - 50 dBA

Shocks

- Spinning hard drives are fragile and sensitive to shocks
- Disk head and platter collisions result in rapid data loss

Key Premise

Non-rotating disks in stand-by mode:

- Consume an order of magnitude less power
- Generate 3x less heat than idle rotating disks
- Are 4-5x less sensitive to shocks

Key Premise

Non-rotating disks in stand-by mode:

- Consume an order of magnitude less power
- Generate 3x less heat than idle rotating disks
- Are 4-5x less sensitive to shocks

Key Premise

Non-rotating disks in stand-by mode:

- Consume an order of magnitude less power
- Generate 3x less heat than idle rotating disks
- Are 4-5x less sensitive to shocks

Key Premise

Non-rotating disks in stand-by mode:

- Consume an order of magnitude less power
- Generate 3x less heat than idle rotating disks
- Are 4-5x less sensitive to shocks

Conclusion: Minimize the time that disks are rotating

Existing Solutions

- Spin the disk up and down
 - Hard to predict future access patterns
 - 3.5" disks have limited number of spin-up cycles (50,000 spin-up cycles max → ~1 per hour)
- Use diskless clients
 - Highly sensitive to network outages
 - High latency
- Use flash-based storage (SSD's)
 - Still quite expensive
 - Sizes available still smaller than enterprises need

Existing Solutions

- Spin the disk up and down
 - Hard to predict future access patterns
 - 3.5" disks have limited number of spin-up cycles (50,000 spin-up cycles max → ~1 per hour)
- Use diskless clients
 - Highly sensitive to network outages
 - High latency
- Use flash-based storage (SSD's)
 - Still quite expensive
 - Sizes available still smaller than enterprises need

Existing Solutions

- Spin the disk up and down
 - Hard to predict future access patterns
 - 3.5" disks have limited number of spin-up cycles (50,000 spin-up cycles max → ~1 per hour)
- Use diskless clients
 - Highly sensitive to network outages
 - High latency
- Use flash-based storage (SSD's)
 - Still quite expensive
 - Sizes available still smaller than enterprises need

Goals

Design a filesystem that:

- Minimizes the time that local system disks are turned on
 - Spin up local hard disks for short periods of time, and only several times/day
- Provides performance comparable to existing filesystems on local hard disks
- Improves reliability through continuous and whole data protection
 - Do so even when systems lose network connectivity
- Requires minimal infrastructure changes and implementation costs

Goals

Design a filesystem that:

- Minimizes the time that local system disks are turned on
 - Spin up local hard disks for short periods of time, and only several times/day
- Provides performance comparable to existing filesystems on local hard disks
- Improves reliability through continuous and whole data protection
 - Do so even when systems lose network connectivity
- Requires minimal infrastructure changes and implementation costs

Goals

Design a filesystem that:

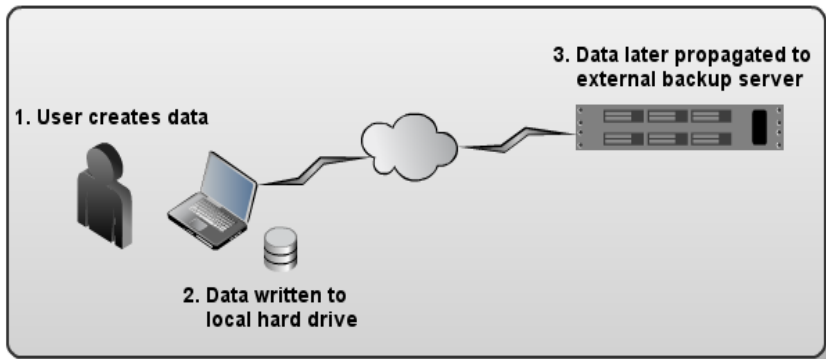
- Minimizes the time that local system disks are turned on
 - Spin up local hard disks for short periods of time, and only several times/day
- Provides performance comparable to existing filesystems on local hard disks
- Improves reliability through continuous and whole data protection
 - Do so even when systems lose network connectivity
- Requires minimal infrastructure changes and implementation costs

Goals

Design a filesystem that:

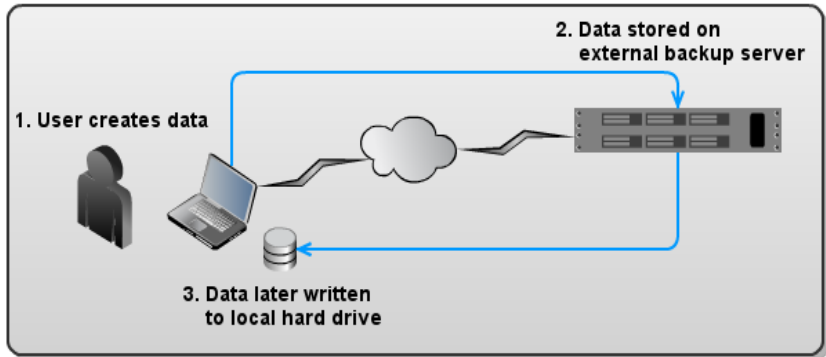
- Minimizes the time that local system disks are turned on
 - Spin up local hard disks for short periods of time, and only several times/day
- Provides performance comparable to existing filesystems on local hard disks
- Improves reliability through continuous and whole data protection
 - Do so even when systems lose network connectivity
- Requires minimal infrastructure changes and implementation costs

Traditional Filesystem / Backup Model



GreenFS Model

GreenFS effectively reverses this model:



Flash Buffer

- Layer of flash memory (e.g. 1 GB USB stick) acts as a buffer between GreenFS and the remote server
- Stores updates before they are propagated to the server
- Caches most recent version of frequently accessed files
- Read latency defined by flash layer and not the network

Flash Buffer

- Layer of flash memory (e.g. 1 GB USB stick) acts as a buffer between GreenFS and the remote server
- Stores updates before they are propagated to the server
- Caches most recent version of frequently accessed files
- Read latency defined by flash layer and not the network

Flash Buffer

- Layer of flash memory (e.g. 1 GB USB stick) acts as a buffer between GreenFS and the remote server
- Stores updates before they are propagated to the server
- Caches most recent version of frequently accessed files
- Read latency defined by flash layer and not the network



Flash Buffer

- Layer of flash memory (e.g. 1 GB USB stick) acts as a buffer between GreenFS and the remote server
- Stores updates before they are propagated to the server
- Caches most recent version of frequently accessed files
- Read latency defined by flash layer and not the network

Local Disk

- Keep disks turned off most of the time
 - Low power consumption
 - Less noise
 - Reduced sensitivity to shock
- Local disk effectively becomes a backup of the data stored remotely
- Use local disk only:
 - When disconnected from the network
 - When network is insufficient for high-bandwidth workloads
- Must periodically synchronize the local disk with the backup server

Local Disk

- Keep disks turned off most of the time
 - Low power consumption
 - Less noise
 - Reduced sensitivity to shock
- Local disk effectively becomes a backup of the data stored remotely
- Use local disk only:
 - When disconnected from the network
 - When network is insufficient for high-bandwidth workloads
- Must periodically synchronize the local disk with the backup server

Local Disk

- Keep disks turned off most of the time
 - Low power consumption
 - Less noise
 - Reduced sensitivity to shock
- Local disk effectively becomes a backup of the data stored remotely
- Use local disk only:
 - When disconnected from the network
 - When network is insufficient for high-bandwidth workloads
- Must periodically synchronize the local disk with the backup server

Local Disk

- Keep disks turned off most of the time
 - Low power consumption
 - Less noise
 - Reduced sensitivity to shock
- Local disk effectively becomes a backup of the data stored remotely
- Use local disk only:
 - When disconnected from the network
 - When network is insufficient for high-bandwidth workloads
- Must periodically synchronize the local disk with the backup server

Local Disk

- Keep disks turned off most of the time
 - Low power consumption
 - Less noise
 - Reduced sensitivity to shock
- Local disk effectively becomes a backup of the data stored remotely
- Use local disk only:
 - When disconnected from the network
 - When network is insufficient for high-bandwidth workloads
- **Must periodically synchronize the local disk with the backup server**

Synchronization

Data updates synchronized with the local disk from the backup server:

- At the time of system shutdown
 - Guarantees data availability even if network is unavailable at next boot
- Periodically at a configurable frequency
 - GreenFS tracks rate of spin-up operations and will not spin-up the local disk if it was spun up too often in the last several days
- At user request

Synchronization

Data updates synchronized with the local disk from the backup server:

- At the time of system shutdown
 - Guarantees data availability even if network is unavailable at next boot
- Periodically at a configurable frequency
 - GreenFS tracks rate of spin-up operations and will not spin-up the local disk if it was spun up too often in the last several days
- At user request

Synchronization

Data updates synchronized with the local disk from the backup server:

- At the time of system shutdown
 - Guarantees data availability even if network is unavailable at next boot
- Periodically at a configurable frequency
 - GreenFS tracks rate of spin-up operations and will not spin-up the local disk if it was spun up too often in the last several days
- At user request

Flash Layer Reliability

- Flash memory has a limited number of write cycles

Type	Max. Overwrites	
Single-Level Cell (SLC)	100,000	Faster, more expensive
Multi-Level Cell (MLC)	10,000	Cheaper, bigger

- USB drives have built-in wear leveling. To wear out a 4 GB drive, with constant writing at full bandwidth, it would take:

Type	Expected Lifetime
SLC	1 year
MLC	1 month

- Under typical workloads, SLC flash should last decades
- Still, it is possible that a worn out cell could result in propagation of corrupted data to the backup server

Flash Layer Reliability

- Flash memory has a limited number of write cycles

Type	Max. Overwrites	
Single-Level Cell (SLC)	100,000	Faster, more expensive
Multi-Level Cell (MLC)	10,000	Cheaper, bigger

- USB drives have built-in wear leveling. To wear out a 4 GB drive, with constant writing at full bandwidth, it would take:

Type	Expected Lifetime
SLC	1 year
MLC	1 month

- Under typical workloads, SLC flash should last decades
- Still, it is possible that a worn out cell could result in propagation of corrupted data to the backup server

Flash Layer Reliability

- Flash memory has a limited number of write cycles

Type	Max. Overwrites	
Single-Level Cell (SLC)	100,000	Faster, more expensive
Multi-Level Cell (MLC)	10,000	Cheaper, bigger

- USB drives have built-in wear leveling. To wear out a 4 GB drive, with constant writing at full bandwidth, it would take:

Type	Expected Lifetime
SLC	1 year
MLC	1 month

- Under typical workloads, SLC flash should last decades
- Still, it is possible that a worn out cell could result in propagation of corrupted data to the backup server

Flash Layer Reliability

- Flash memory has a limited number of write cycles

Type	Max. Overwrites	
Single-Level Cell (SLC)	100,000	Faster, more expensive
Multi-Level Cell (MLC)	10,000	Cheaper, bigger

- USB drives have built-in wear leveling. To wear out a 4 GB drive, with constant writing at full bandwidth, it would take:

Type	Expected Lifetime
SLC	1 year
MLC	1 month

- Under typical workloads, SLC flash should last decades
- Still, it is possible that a worn out cell could result in propagation of corrupted data to the backup server

Flash Layer Reliability

- To compensate, versioning is used on the backup server
- If a corrupted update is sent to the backup server, the user can always revert to a previous version of a file
- Thus, it should be safe to use cheaper MLC flash

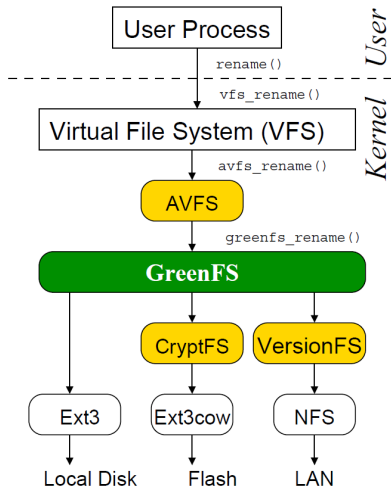
Flash Layer Reliability

- To compensate, versioning is used on the backup server
- If a corrupted update is sent to the backup server, the user can always revert to a previous version of a file
- Thus, it should be safe to use cheaper MLC flash

Flash Layer Reliability

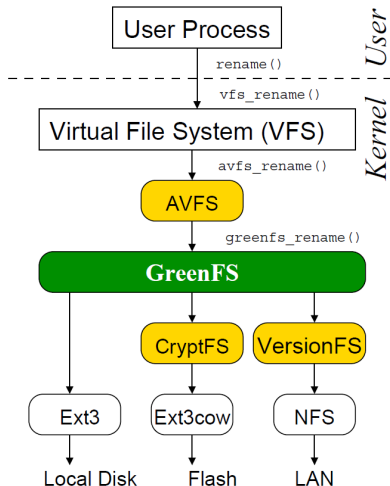
- To compensate, versioning is used on the backup server
- If a corrupted update is sent to the backup server, the user can always revert to a previous version of a file
- Thus, it should be safe to use cheaper MLC flash

Implementation



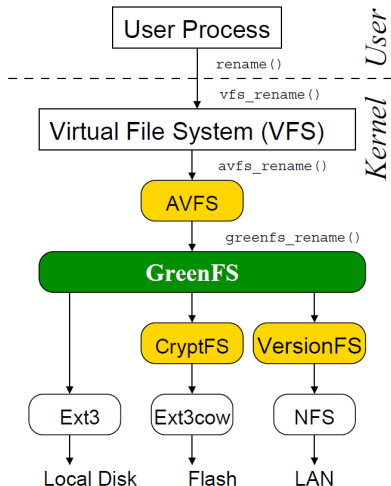
- Implemented as a loadable Linux kernel module
- Fan-out stackable filesystem
- Can add features as needed by stacking:
 - Encryption
 - Anti-virus protection

Implementation



- Implemented as a loadable Linux kernel module
- Fan-out stackable filesystem
- Can add features as needed by stacking:
 - Encryption
 - Anti-virus protection

Implementation



- Implemented as a loadable Linux kernel module
- Fan-out stackable filesystem
- Can add features as needed by stacking:
 - Encryption
 - Anti-virus protection

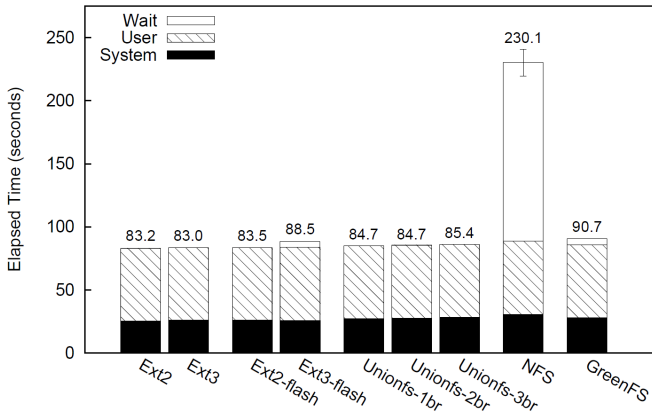
Test Environment

Backup Server	IBM LS20 Blade	2 SCSI 10KRPM disks
Server	IBM xSeries 225	1 SCSI 10KRPM disk
Desktop	IBM ThinkCentre 8187	1 IDE disk
Notebook	IBM ThinkPad T42	1 IDE disk

- 1 GB MLC flash drives used for each flash layer
- 1 Gbps network connecting systems

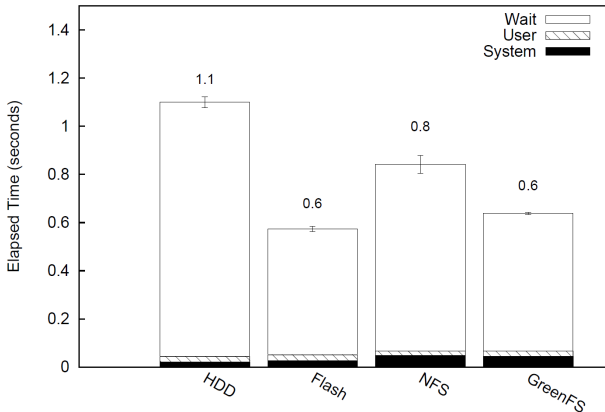
Benchmark 1: OpenSSH Compilation

- CPU-bound; generates a lot of writes



Benchmark 2: Emacs Invocation

- Local disk slowest: disk heads take time to reposition



Power Efficiency

- Idle systems:

System	Original Power (W)	GreenFS Power (W)	Savings (%)
Blade Server	90	n/a	n/a
Desktop	113	101	12
Notebook	54.1	46.8	14
Notebook 2	20.1	19.6	2.5
	13.9	13.4	3.6

- Active Systems:

- Workloads executed based on traces from normal kernel developer activity
- Up to 60% storage power savings realized

Shock Protection

- Notebook taken up several flights of stairs and on an elevator
- GreenFS compared with Active Protection System (APS)
 - Parks HD heads when it detects a laptop may be falling
 - Does not protect against external impacts

	Elevator	Stairs
Trip time (s)	85	58
Network available (%)	94	100
APS shock protection (%)	29	63
GreenFS shock protection (%)	100	100
APS disk stop operations	4	6
GreenFS disk stop operations	0	0

Shock Protection

- Notebook taken up several flights of stairs and on an elevator
- GreenFS compared with Active Protection System (APS)
 - Parks HD heads when it detects a laptop may be falling
 - Does not protect against external impacts

	Elevator	Stairs
Trip time (s)	85	58
Network available (%)	94	100
APS shock protection (%)	29	63
GreenFS shock protection (%)	100	100
APS disk stop operations	4	6
GreenFS disk stop operations	0	0

Summary

- GreenFS reduces power, heat, and noise by keeping local disks turned off
- Reduces sensitivity of disks to shocks, including external impacts
- Performance is equal or superior to existing filesystems
- Reliability is enhanced through continuous, whole-data protection
- Requires few changes to existing infrastructure

Summary

- GreenFS reduces power, heat, and noise by keeping local disks turned off
- Reduces sensitivity of disks to shocks, including external impacts
- Performance is equal or superior to existing filesystems
- Reliability is enhanced through continuous, whole-data protection
- Requires few changes to existing infrastructure

Summary

- GreenFS reduces power, heat, and noise by keeping local disks turned off
- Reduces sensitivity of disks to shocks, including external impacts
- Performance is equal or superior to existing filesystems
- Reliability is enhanced through continuous, whole-data protection
- Requires few changes to existing infrastructure

Summary

- GreenFS reduces power, heat, and noise by keeping local disks turned off
- Reduces sensitivity of disks to shocks, including external impacts
- Performance is equal or superior to existing filesystems
- Reliability is enhanced through continuous, whole-data protection
- Requires few changes to existing infrastructure

Summary

- GreenFS reduces power, heat, and noise by keeping local disks turned off
- Reduces sensitivity of disks to shocks, including external impacts
- Performance is equal or superior to existing filesystems
- Reliability is enhanced through continuous, whole-data protection
- Requires few changes to existing infrastructure

Questions?