NOVA: The Fastest File System for NVDIMMs

Steven Swanson, UC San Diego

XFS F2FS NILFS

EXT4 BTRFS

Disk-based file systems are inadequate for NVMM

- Disk-based file systems cannot exploit NVMM performance
- Performance optimization compromises consistency on system failure [1]

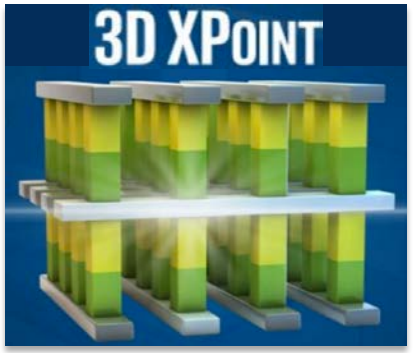
Atomicity	1-Sector overwrite	1-Sector append	1-Block overwrite	1-Block append	N-Block overwrite	N-Block append
Ext4 wb	✓	✗	✗	✗	✗	✗
Ext4 Order	✓	✓	✗	✓	✗	✓
Ext4 Dataj	✓	✓	✓	✓	✗	✓
Btrfs	✓	✓	✓	✓	✗	✓
xfs	✓	✓	✗	✓	✗	✓
Reiserfs	✓	✓	✗	✓	✗	✓

[1] Pillai *et al*, All File Systems Are Not Created Equal: On the Complexity of Crafting Crash-Consistent Applications, OSDI '14.

BPFS SCMFS PMFS Aerie

EXT4-DAX XFS-DAX

NOVA M1FS



Previous Prototype NVMM file systems are not strongly consistent

- DAX does not provide data atomicity guarantee
- So programming is more difficult

Atomicity	Metadata	Data	ATomic Mmap [1]	Snapshot
BPFS	✓	✓ [2]	✗	✗
PMFS	✓	✗	✗	✗
Ext4-DAX	✓	✗	✗	✗
Xfs-DAX	✓	✗	✗	✗
SCMFS	✗	✗	✗	✗
Aerie	✓	✗	✗	✗

Ext4-DAX and xfs-DAX shortcomings

- No data atomicity support
- Single journal shared by all the transactions (JBD2-based)
- Poor performance
- Development teams are (rightly) “disk first”.

NOVA provides strong atomicity guarantees

Atomicity	1-Sector overwrite	1-Sector append	1-Block overwrite	1-Block append	N-Block overwrite	N-Block append
Ext4 wb	✓	✗	✗	✗	✗	✗
Ext4 Order	✓	✓	✗	✓	✗	✓
Ext4 Dataj	✓	✓	✓	✓	✗	✓
Btrfs	✓	✓	✓	✓	✗	✓
xfs	✓	✓	✗	✓	✗	✓
Reiserfs	✓	✓	✗	✓	✗	✓

Atomicity	Metadata	Data	Mmap
BPFS	✓	✓	✗
PMFS	✓	✗	✗
Ext4-DAX	✓	✗	✗
Xfs-DAX	✓	✗	✗
SCMFS	✗	✗	✗
Aerie	✓	✗	✗

NOVA provides strong atomicity guarantees

Atomicity	1-Sector overwrite	1-Sector append	1-Block overwrite	1-Block append	N-Block overwrite	N-Block append
Ext4 wb	✓	✗	✗	✗	✗	✗
Ext4 Order	✓	✓	✗	✓	✗	✓
Ext4 Dataj	✓	✓	✓	✓	✗	✓
Btrfs	✓	✓	✓	✓	✗	✓
xfs	✓	✓	✗	✓	✗	✓
Reiserfs	✓	✓	✗	✓	✗	✓
NOVA	✓	✓	✓	✓	✓	✓

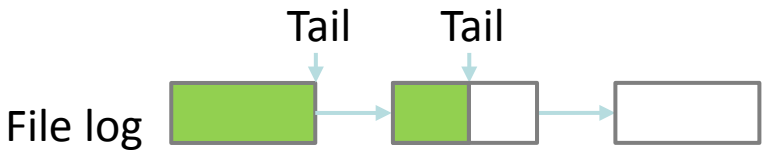
Atomicity	Metadata	Data	Mmap
BPFS	✓	✓	✗
PMFS	✓	✗	✗
Ext4-DAX	✓	✗	✗
Xfs-DAX	✓	✗	✗
SCMFS	✗	✗	✗
Aerie	✓	✗	✗
NOVA	✓	✓	✓

NOVA: A File System for NVDIMMs

- Dispenses with disk-centric baggage
  - Strong consistency guarantees
  - Lean and fast
- Freely optimize for NVDIMMs
  - No need to worry about future disk-based features or optimizations.
  - E.g., Userspace f/msync() is no problem.
- Focus on latency and bandwidth advantages of NVDIMMs
- Freely available (GPL)

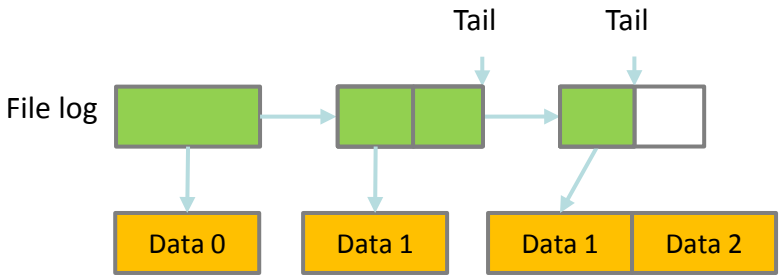
Log Structure + copy-on-write + Journals

- One log per iNode
- Non-contiguous
- Fast, Simple atomic updates
- Meta-data only

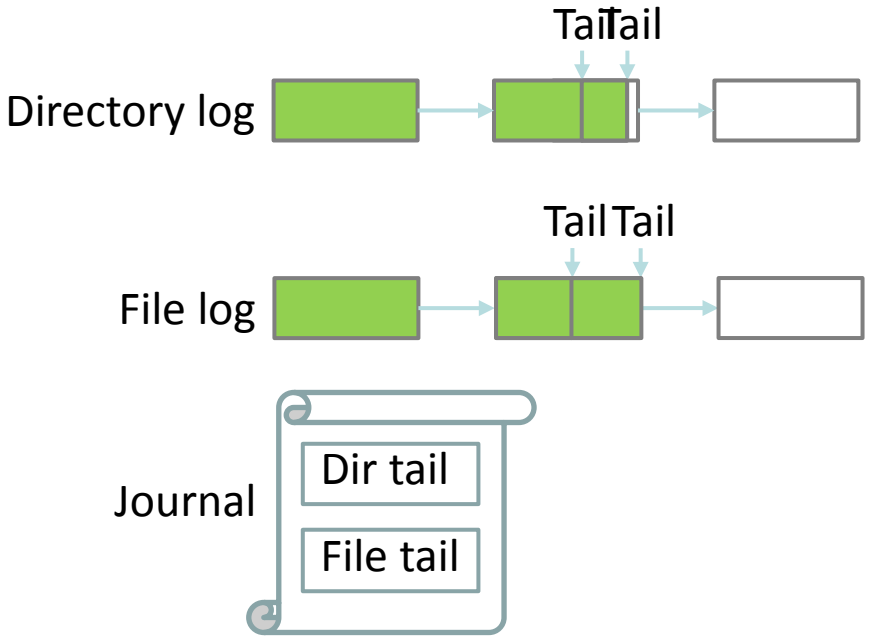


Log Structure + copy-on-write + Journals

- Multi-page atomic update
- Fast allocation
- Instant data GC



Log Structure + copy-on-write + Journals

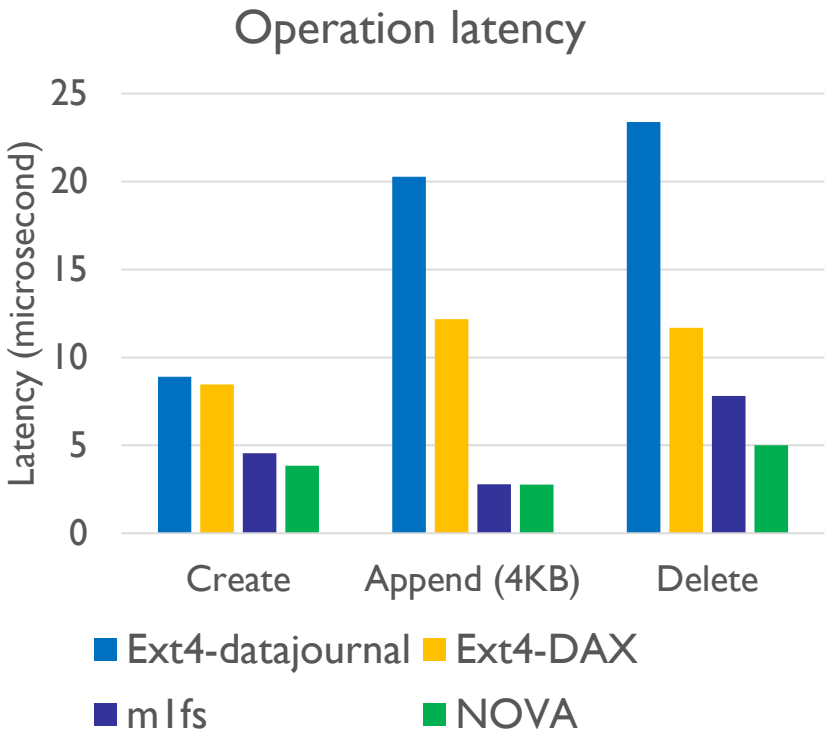


- Small, fixed sized journals
- For complex ops.

NOVA Reliability Features

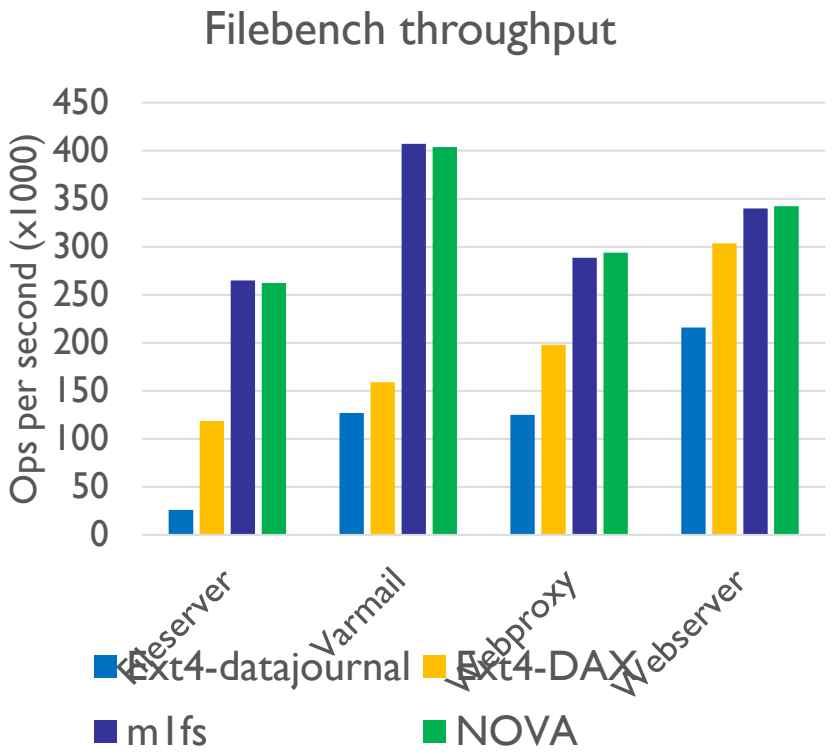
- NVDIMM file systems are subject to multiple failure modes
  - Media errors – “hard” and “soft”
  - Local system failures
  - Large scale failures
- NOVA includes reliability mechanisms (April release)
  - Replicated metadata
  - CRC for metadata
  - Erasure codes for file data
  - Read-only checkpoints for backup
- NOVA includes “unsafe” modes that improve performance but sacrifice consistency (April release)

Evaluation: Latency



- Intel PM Emulation Platform
  - Emulates different NVM characteristics
  - Emulates clwb/PCOMMIT latency
- NOVA provides low latency atomicity

Filebench throughput

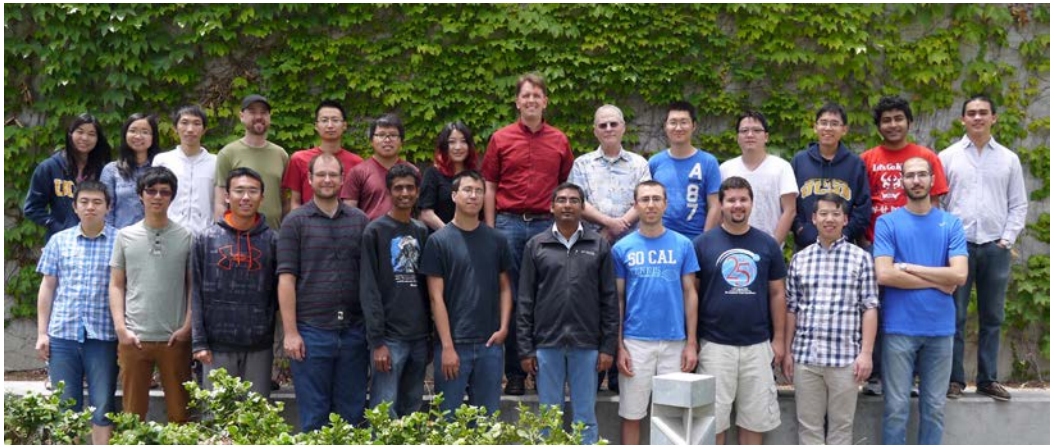


- NOVA achieves high performance with strong data consistency

Conclusion

- Existing file systems do not meet the requirements of applications on NVMM file systems
- NOVA's multi-log design achieves high concurrency, efficient garbage collection and fast recovery
- NOVA outperforms existing file systems while providing stronger consistency and atomicity guarantees
- NOVA protects data against media errors and system failures.

Thank you!



Try NOVA!

https://github.com/NVSL/NOVA