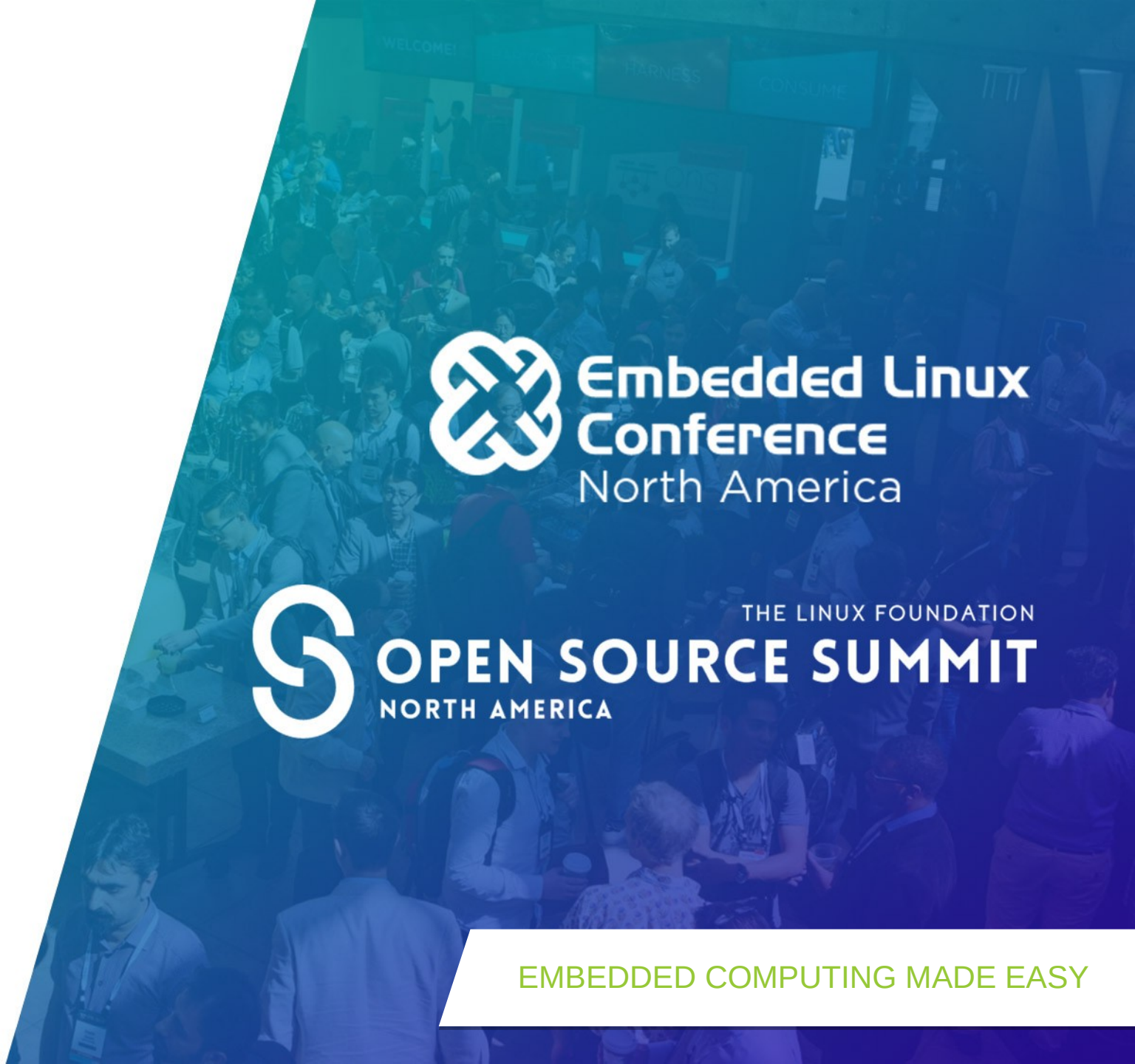
Toradex
Embedded. Computing.

Wear Estimation for Devices with eMMC Flash Memory



EMBEDDED COMPUTING MADE EASY

WITH YOU TODAY…

- Joined Toradex 2011
- Spearheaded Embedded Linux Adoption
- Introduced Upstream First Policy
- Top 10 U-Boot Contributor
- Top 10 Linux Kernel ARM SoC Contributor
- Industrial Embedded Linux Platform Torizon Fully Based on Mainline Technology
  - Mainline U-Boot with Distroboot
  - KMS/DRM Graphics with Etnaviv & Nouveau
  - OTA with OSTree
  - Docker

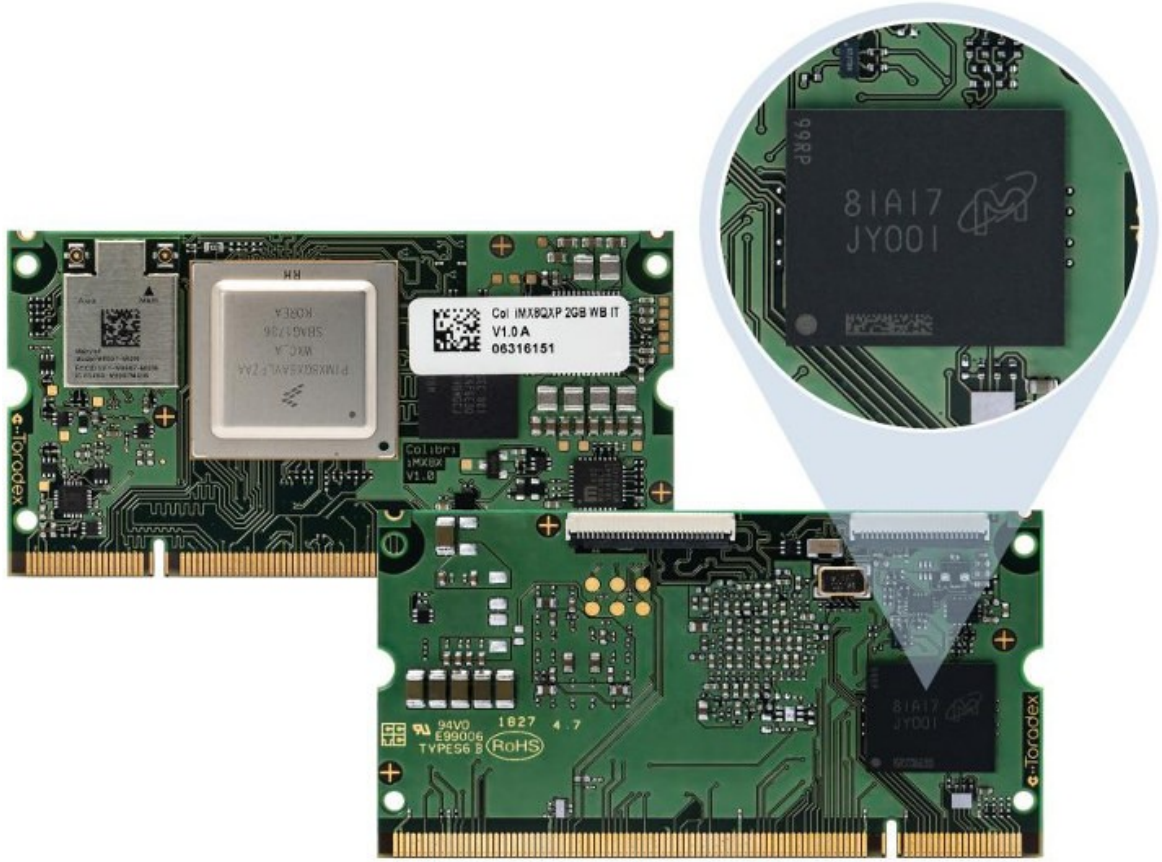
Marcel Ziswiler
Platform Manager Embedded Linux
marcel.ziswiler@toradex.com
Toradex AG

WHAT WE'LL COVER TODAY

- A Technology Overview
- eMMC
- Flash Health
- I/O Tracking
- Lifespan Estimation
- Flash Analytics Tool
- Conclusion

Toradex
Embedded. Computing.

Flash – Non-Volatile Memory of Choice

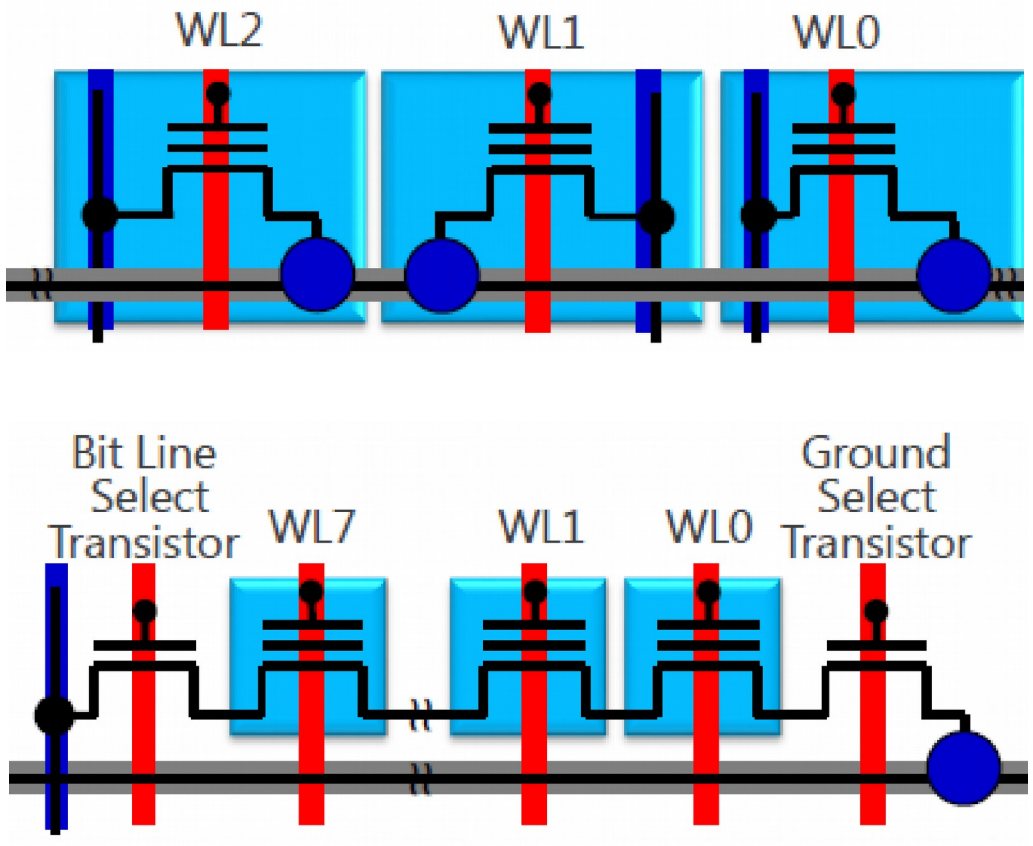


In Embedded Systems

- Decreased Size
- Increased Robustness
- No Moving Parts
- Reduced Power Consumption
- Keep Redundant Data On-Site
- For Intermittent Connectivity Reasons

Toradex
Embedded. Computing.

NOR vs. NAND



- Difference at Transistor Level How to Store Bits
- NOR and NAND Logic Gates

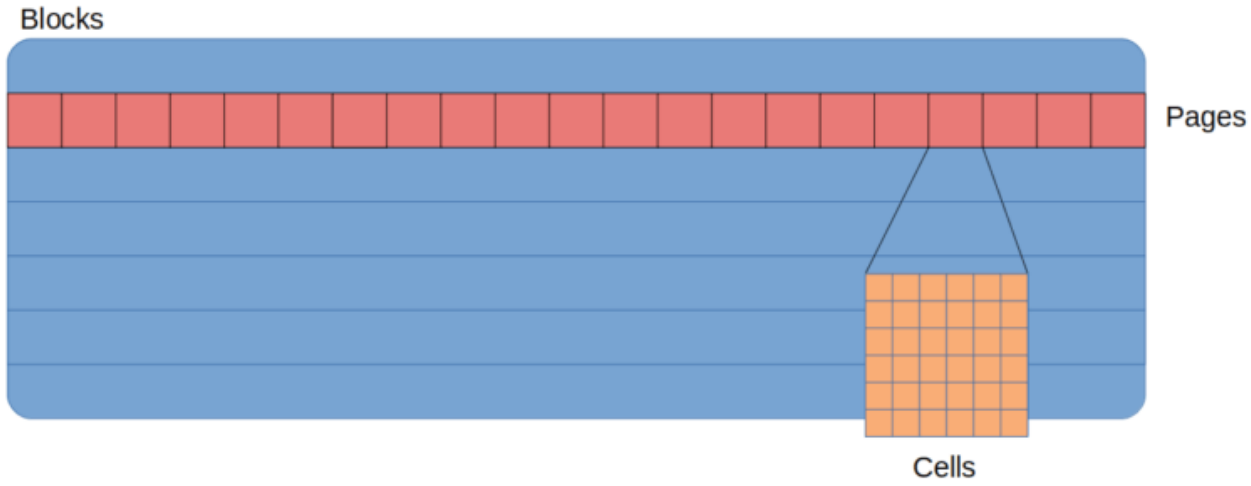
- Simpler Principle of Operation
- Higher Reliability
- Higher Pin-Count
- Lower Density in Silicon
- Bigger Size
- More Expensive
- Only for Specific Applications
- Highly Critical Industrial-Grade

Toradex
Embedded. Computing.

NAND Structure

Cell

- Smallest Entity
- Storing Data at Bit-Level



Page

- Smallest Array of Cells
- Addressable for Read/Write Operations
- Flipping Bits from 1 to 0
- Page Size: Range of Kilobytes e.g. 4 kB

(Erase-)Block

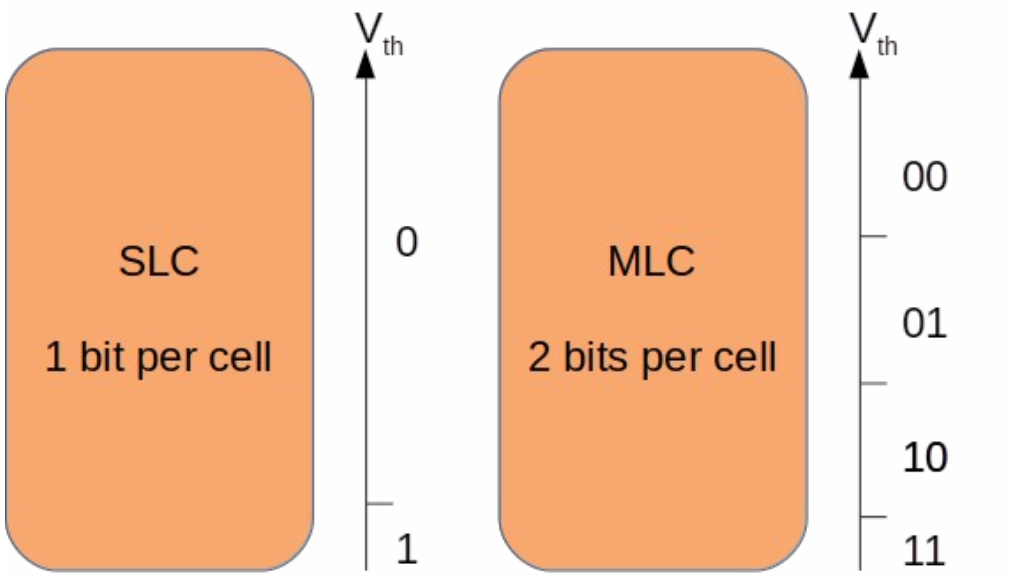
- Smallest Array of Pages
- Addressable for Erase Operation
- Return Logic State of Bits from 0 Back to 1
- Block Size: Range of Megabytes e.g. 4 MB
- Erase Operation is Slow
- Wears out Flash over Time
- Develops Bad Blocks
- Block Erase Count

Toradex
Embedded. Computing.

NAND: SLC vs. MLC

Cell

- How Many Bits Stored
- Depends on Voltage Level Thresholds



SLC

- Single-Level Cell
- Stores 1 Bit per Cell

pSLC

- Pseudo-SLC
- MLC Operating in SLC Mode
- Stores 1 Bit per Cell

MLC

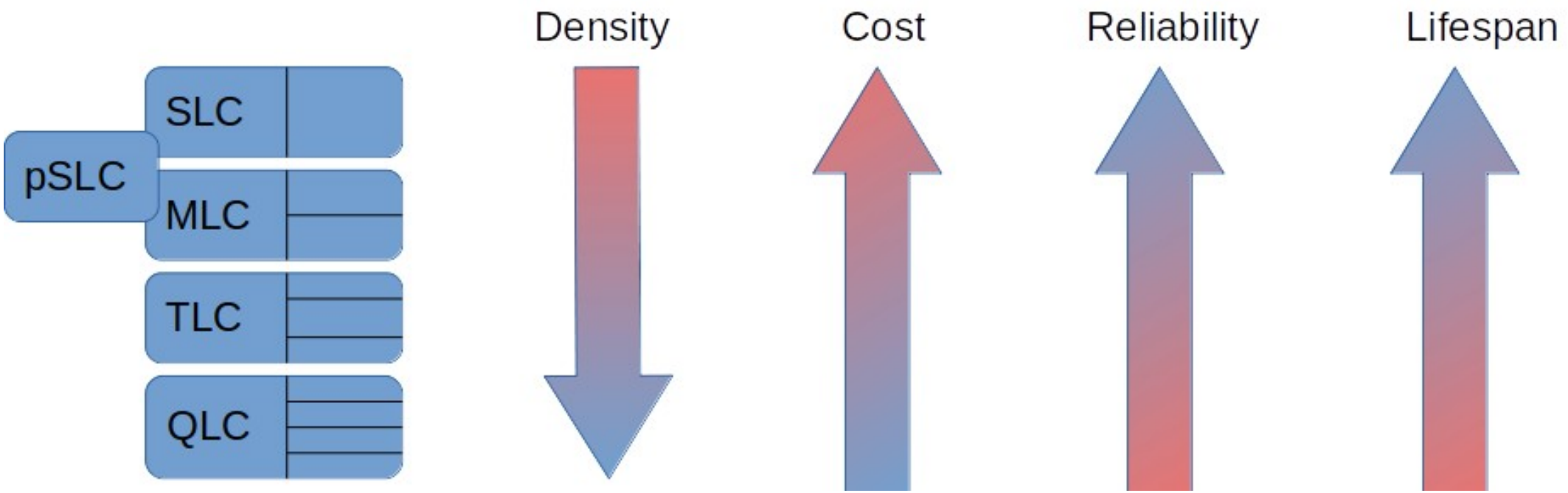
- Multi-Level Cell
- Stores 2 Bits per Cell

TLC, QLC, ...

- You Get the Idea...

Toradex
Embedded. Computing.

Trade-Off Between Density and Cost vs. Reliability and Lifespan



Toradex
Embedded. Computing.

ECC and Bad Blocks

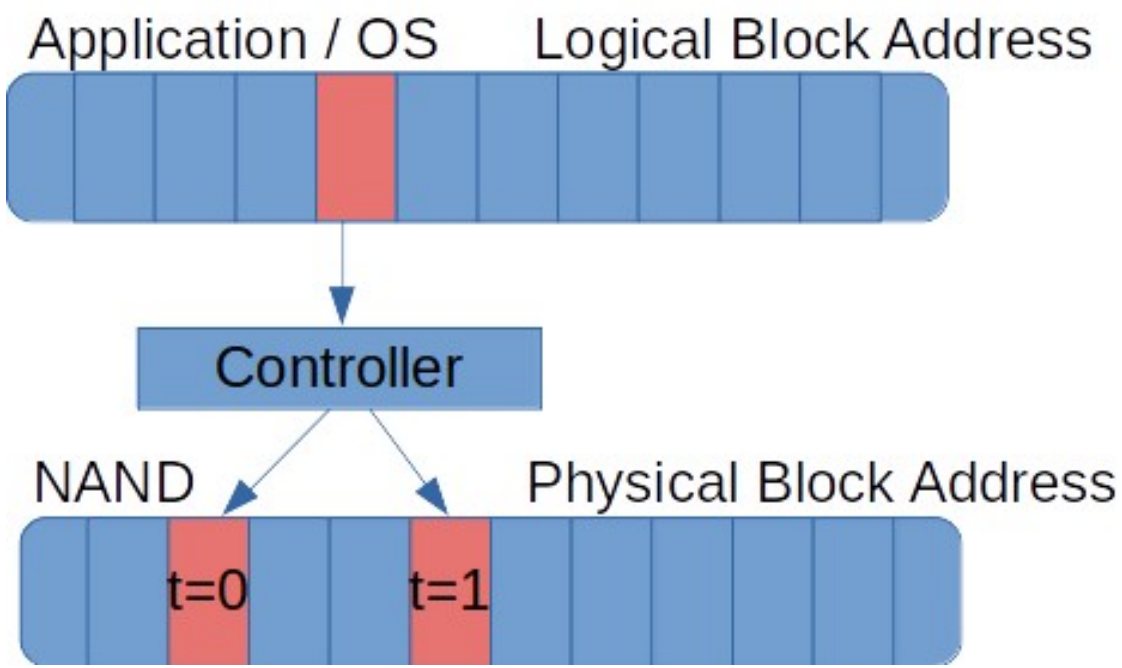
Error Correction Code Algorithms

- Adding Redundancy
- Allow Correcting resp. Detecting Certain Bit Errors
- Random Bit Flips Even in Healthy Blocks

Bad Blocks

- Over Time Probability of Bit-Flips Increases
- Blocks Wear out Becoming Bad
- Factory Bad Blocks
- Spare Blocks

Wear-Leveling and Garbage Collection



Wear-Leveling

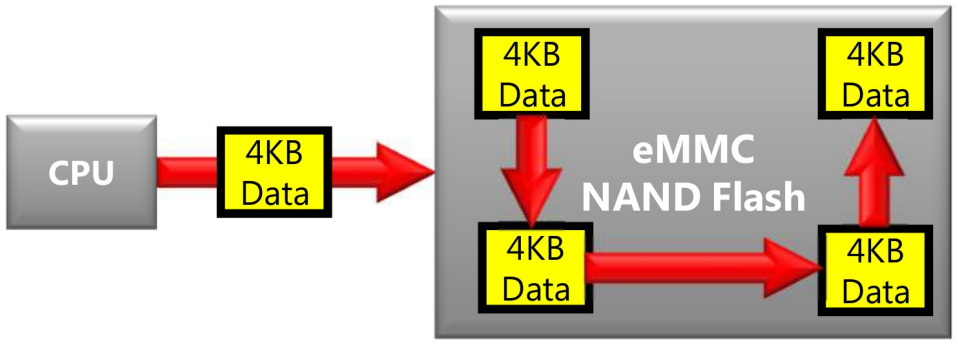
- Same Physical Pages/Blocks Used for e.g. File Update
- Increased Wear out Causing Premature Bad Blocks
- Using Blocks Evenly
- Moving Data Around
- Dynamic vs. Static

Garbage Collection

- Slow Erase Operation
- Avoid Immediate Erasure
- Just Marking Blocks Dirty
- Erase Later e.g. Idle Time

Toradex
Embedded. Computing.

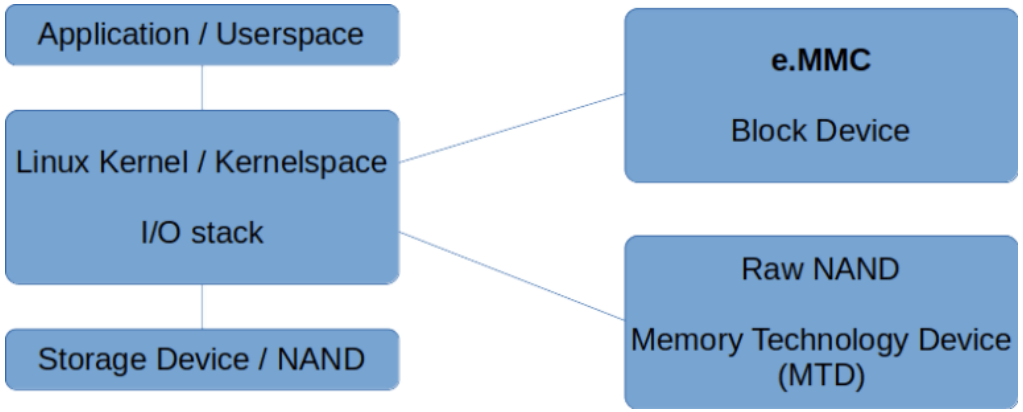
Write Amplification Factor (WAF)



- Actual Data Written to NAND Flash Cells vs.
- Data Sent from Host to Memory
- Difference Between Programm and Erase Size
- Data Needs Erasing Before (Re-)Writing
- Memory Management Features:
  - Wear-Leveling
  - Garbage Collection
- Typical WAF in eMMC: Good Average is 4
- Depends on Usage Scenario
- Select Optimal Data Size Related to Page Size

Toradex
Embedded. Computing.

Embedded MultiMediaCard (eMMC)



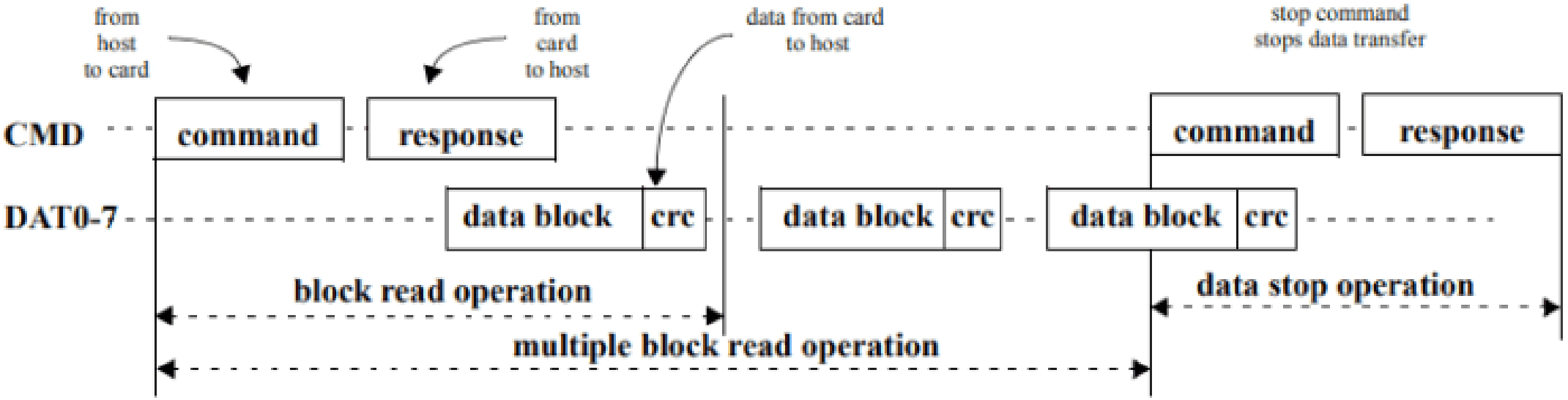
Managed NAND

- Raw NAND Die & Accompanying NAND Controller
- Abstracting Large Part of Management SW-Stack
- Latest JEDEC Standard 5.1
- Allows for Regular Block Device Operations
- Using Regular File Systems e.g. EXT4
- Example eMMC
  - Micron MTFC4GACAJCN-1M-WT
  - 4 GB MLC
  - 1024 Blocks of 4 KB Size
  - Lifespan 3000 Write/Erase Cycles
  - 15 nm Process

Toradex
Embedded. Computing.

MMC Protocol

- Bus: Command, Clock and 7 Data Lines
- CMD: Serial Command/Response Channel
- DAT0-7: Parallel Read/Write Data plus CRC
- Single or Multiple Block Read/Write Operations



Toradex
Embedded. Computing.

MMC Registers

Name	*Width (bytes)*	*Description*
CID	16	Unique Card/Device Identifier.
RCA	2	Relative Card/Device Address: device's system address, dynamically assigned by the host during initialization.
DSR	2	Driver Stage Register: to configure the device's output drivers.
CSD	16	Card/Device Specific Data: information about the device's operation conditions.
OCR	4	Operation Conditions Register: used by a special broadcast command to identify the voltage type of the device.
EXT_CSD	512	Extended Card/Device Specific Data: contains information about the device's capabilities and selected modes. Introduced in standard v4.0.

JEDEC Standard Health Reporting

- Device Life Time Estimation Type A:
  - Health Status in Increments of 10 %
  - Refers to pSLC Blocks in our eMMC
- Device Life Time Estimation Type B:
  - Health Status in Increments of 10 %
  - Refers to MLC Blocks in our eMMC
- Pre-EOL Information:
  - Normal: Up to 80 % of Reserved Blocks Consumed
  - Warning: More than 80 % Consumed
  - Urgent: More than 90 % Consumed
- Introduced with Standard v5.0
- Low Resolution Requiring Very Long Benchmark Runs

Toradex
Embedded. Computing.

Micron Proprietary Health Report

- TN-FC-32: e.MMC Device Health Report
- Bad Block Counters and Information:
  - Factory Bad Block Count
  - Run-Time Bad Block Count
  - Remaining Spare Block Count
  - Per Block Failed Erase vs. Program Operations with Page Addresses
- Block Erase Counters:
  - Minimum, Maximum and Average Among all Blocks
  - Per Block Erase Count
- Block Configuration:
  - Physical Address of Each Block
  - pSLC vs. MLC Configuration
- Accessed by General Command (GEN_CMD) aka CMD56

Flash Health

- Percentage of Capacity Already Worn Out

$$endurance = number\ of\ blocks \cdot average\ block\ lifespan$$

$$endurance = 1024 \cdot 3000 = 3.072.000\ block\ erases$$

or

$$endurance = block\ size \cdot blocks \cdot average\ block\ lifespan$$

$$endurance = 4\ MB \cdot 1024 \cdot 3000 = 12\ TB\ written$$

Toradex
Embedded. Computing.

Monitoring Flash Health in Linux

```
root@colibri-imx6:~# mmc
 Usage:

 mmc extcsd read <device>
 Print extcsd data from <device>.

 mmc extcsd dump <device>
 Print raw extcsd data from <device>.
```

mmc-utils

- Software to Extracts Meaningful Information From eMMC Devices
- Reading Data From Extended Card/Device Specific Data (EXT_CSD)
- Includes Device Lifespan Defined by JEDEC eMMC 5.0 Standard

```
root@colibri-imx6-05097264:/app# mmc extcsd read /dev/mmcblk1
 ===============================================
   Extended CSD rev 1.7 (MMC 5.0)
 ===============================================
```

```
root@colibri-imx6:~# mmc extcsd read /dev/mmcblk1 | grep LIFE
Device life time estimation type B [DEVICE_LIFE_TIME_EST_TYP_B: 0x01]
Device life time estimation type A [DEVICE_LIFE_TIME_EST_TYP_A: 0x01]
eMMC Life Time Estimation A [EXT_CSD_DEVICE_LIFE_TIME_EST_TYP_A]: 0x01
eMMC Life Time Estimation B [EXT_CSD_DEVICE_LIFE_TIME_EST_TYP_B]: 0x01

root@colibri-imx6-05097264:~# mmc extcsd read /dev/mmcblk1 | grep EOL
Pre EOL information [PRE_EOL_INFO: 0x01]
eMMC Pre EOL information [EXT_CSD_PRE_EOL_INFO]: 0x01
```

Toradex
Embedded. Computing.

Vendor Proprietary Health Report

```
/ Retrieve the erase count for each block
// A two-step approach is needed (read number of tables and then read tables)
int do_block_erase_info(int nargs, char **argv)
{
 ret = CMD56_data_in(fd, cmd56_how_many_tables, data_in);
 printf("Block erase count\n");
 printf("Block\tErase\n");
 for(table_idx = 0; table_idx < how_many_tables; table_idx++){
 ret = CMD56_data_in(fd, (table_idx * 256) + cmd56_retrieve_base, data_in);

 for(physical_block = 0; physical_block < 128; physical_block++){
 printf("%d\t%d\n",
         (256*data_in[0+2*physical_block]) + data_in[1+2*physical_block],
         (256*data_in[256+2*physical_block]) + data_in[257+2*physical_block]);
 }
 }
}
```

```
int do_bad_block_count(int nargs, char **argv);
int do_bad_block_info(int nargs, char **argv);
int do_block_erase_count(int nargs, char **argv);
int do_block_erase_info(int nargs, char **argv);
int do_block_addr_type_info(int nargs, char **argv);
```

Toradex
Embedded. Computing.

Vendor Proprietary Health Report 2nd

- Vendor-Specific Tool
- Micron’s emmcparm
- Provides Consolidated Lifespan Report
- More Granular Parameters

```
root@colibri-imx6:~# emmcparm_arm
--spare_block
--bad_block
--erase_count
--sect_count
```

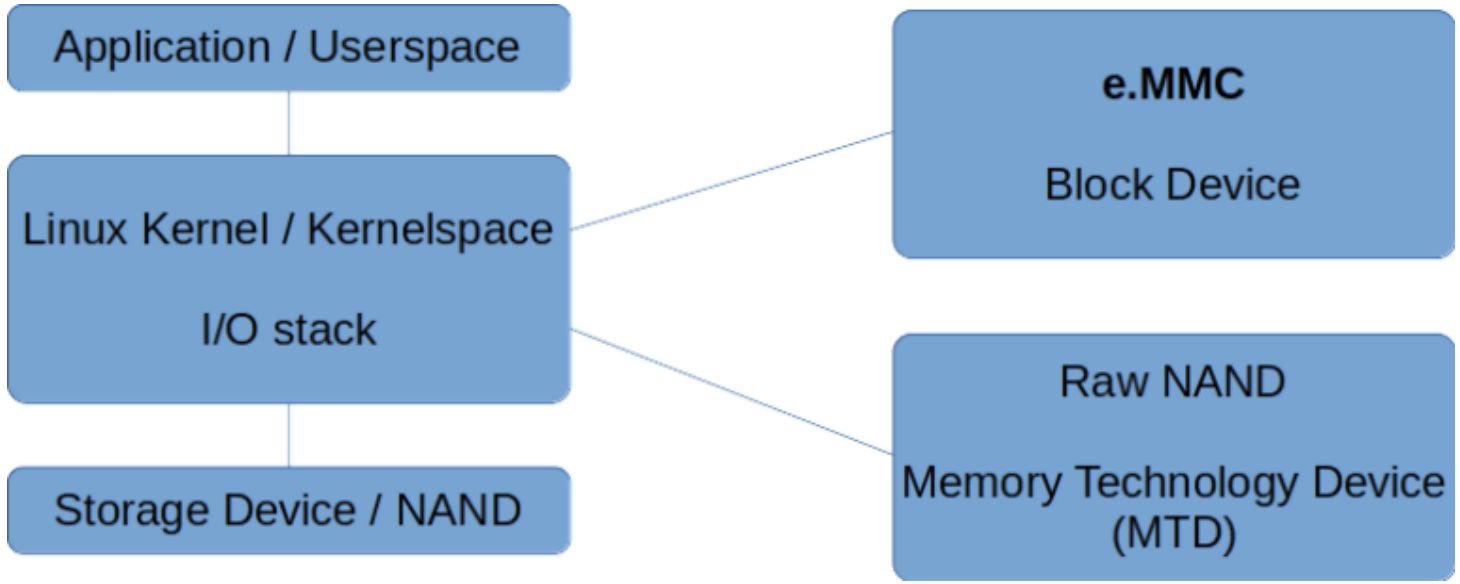
Toradex
Embedded. Computing.

I/O Tracking

- Useful Indicator that Flash Wears out Quickly
- Debug Indicator Showing What Applications Write too Much Data
- Generates Input Data for Wear Estimation Model
- Independent of JEDEC Standards or eMMC Vendor Health Reports
- Applicable to any NAND Flash Based Storage Technology

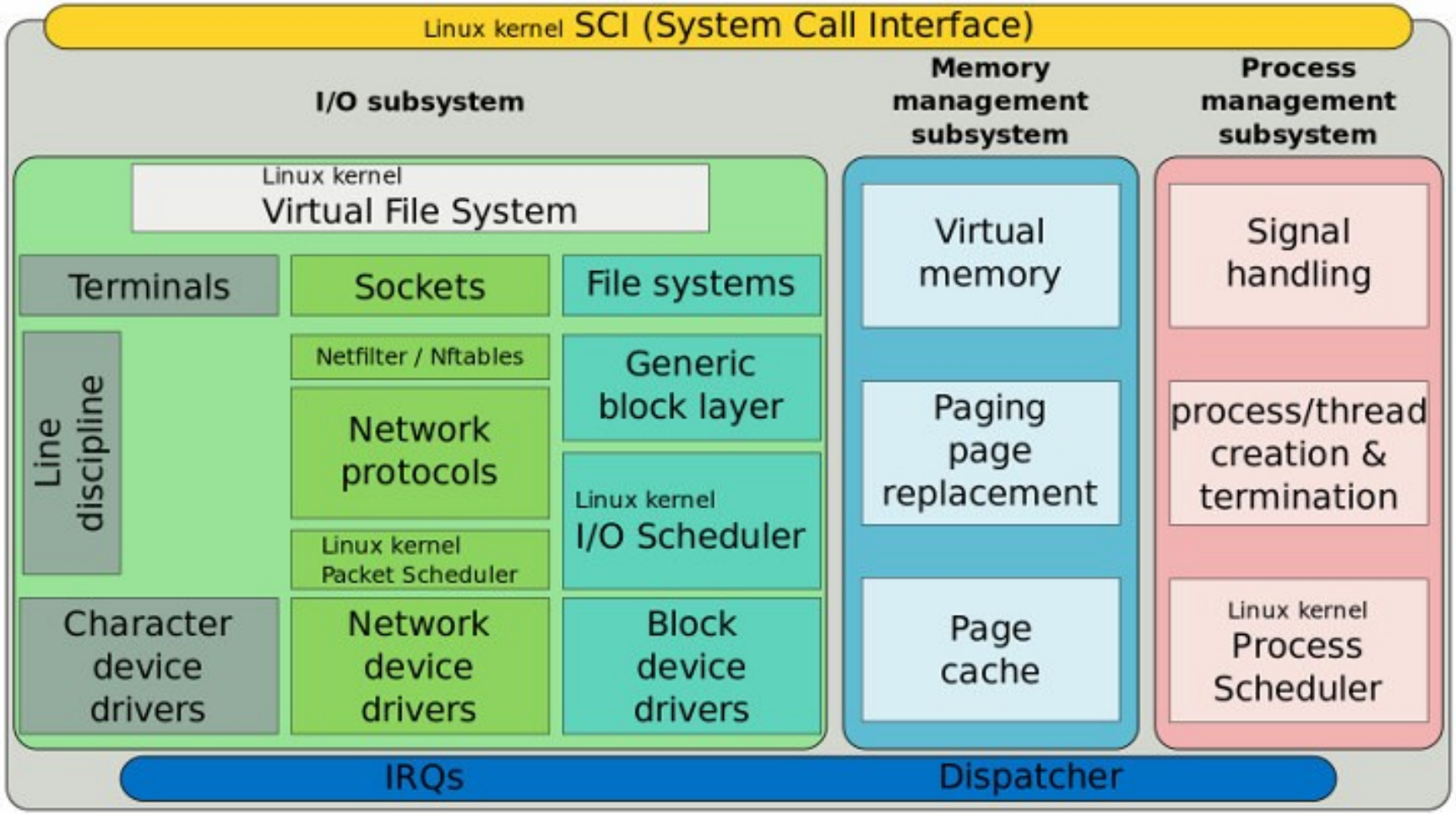
Linux I/O Stack for eMMC and Raw NAND

- Userspace File Operations at Application-Level
- System Calls into Kernelspace
- Ends up in Linux I/O Stack
- Finally Sending Data to Low-Level Device Driver



Toradex
Embedded. Computing.

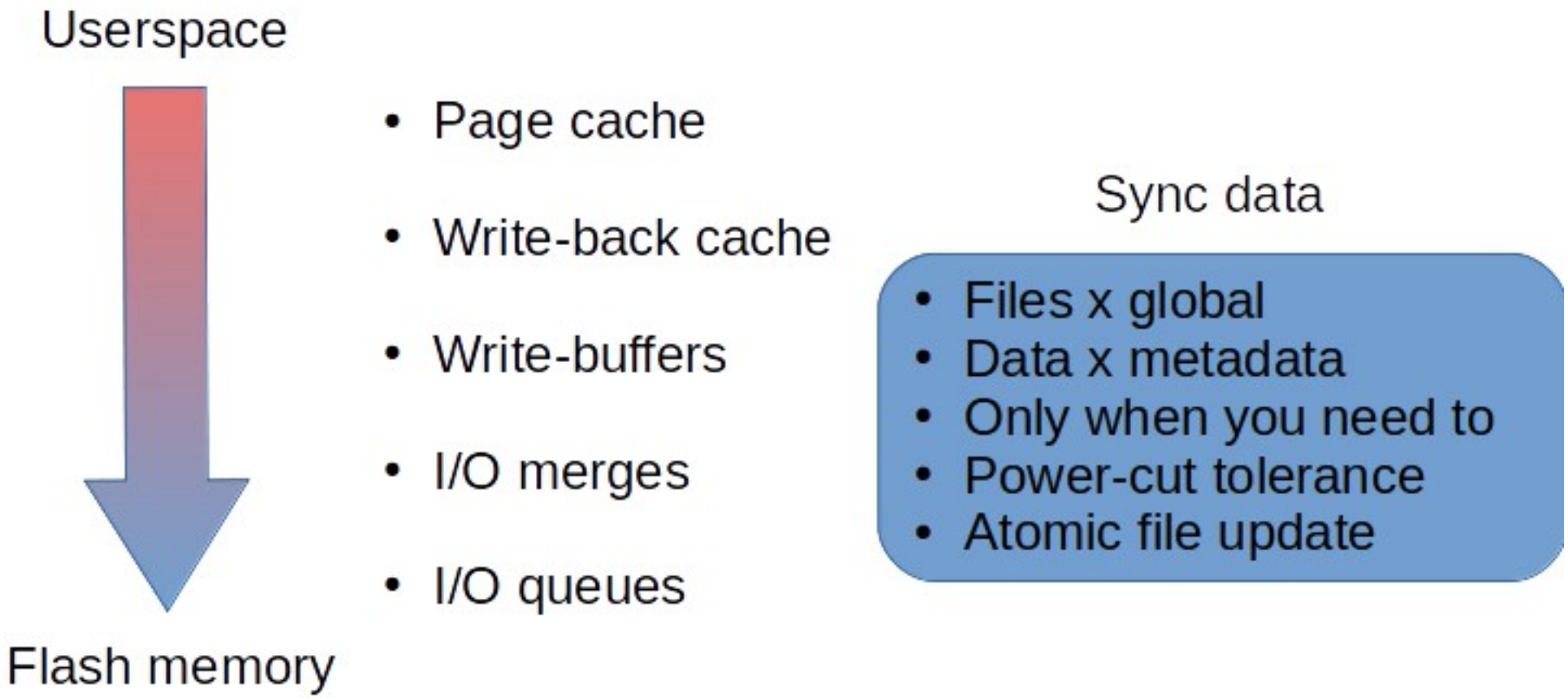
Block Device I/O Stack



- VFS Abstracting Userspace API
- FS File Concept
- Gen Block Layer Handling Block IO
- IO Scheduler Queuing IO Requests
- Max. Block IO Performance

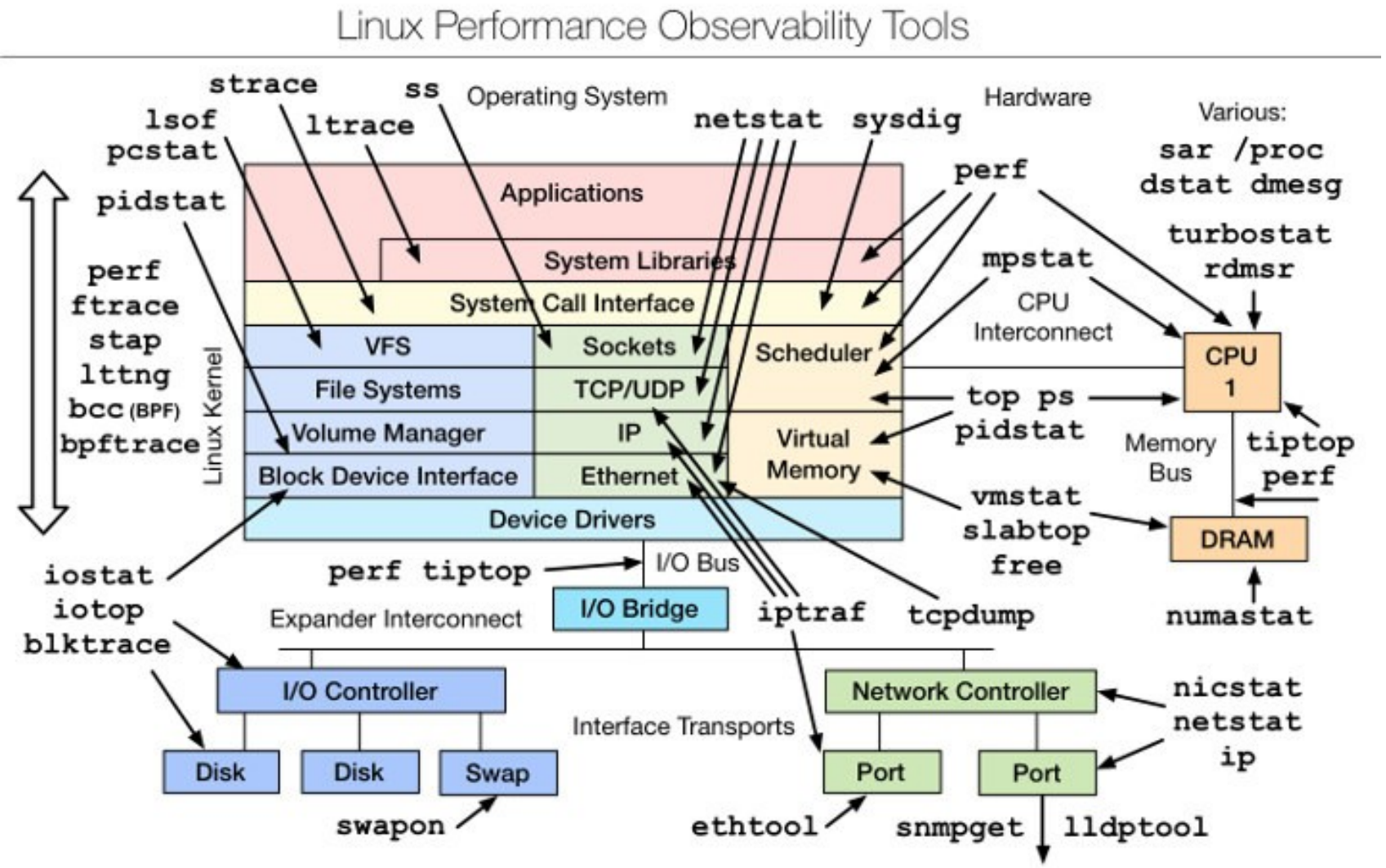
- Why Not Monitor Userspace?
- Not Very Accurate
- Layers of Caches

Caches, Buffers, Queues and Syncs



Toradex
Embedded. Computing.

Measuring I/O Writes



- Monitoring Writes That Actually Hit the Flash
- Where Exactly in the Linux I/O Stack to Measure?
- How to Measure (e.g. What Tool to Use)?

iotop

- Tracking Userspace Operations
- Easy to Use

```
root@colibri-imx6:~# iotop -help
Options:
 -o, --only            only show processes or threads actually doing I/O
 -b, --batch           non-interactive mode
 -a, --accumulated     show accumulated I/O instead of bandwidth
 -k, --kilobytes       use kilobytes instead of a human friendly unit
 -t, --time            add a timestamp on each line (implies -batch)
 -q, --quiet           suppress some lines of header (implies --batch)
```

```
root@colibri-imx6:~# dd if=/dev/urandom bs=4k count=100000 | pv -L 25k > testfile

root@colibri-imx6:~# iotop --only --batch --accumulated --kilobytes --time -quiet
TIME        TID          PRIO    USER       DISK READ  DISK WRITE  SWAPIN      IO    COMMAND
2019-08-02 03:11:19    50 be/4 root          0.00 K     24.00 K -0.00 % -0.00 % pv -L 25k
2019-08-02 03:11:20    50 be/4 root          0.00 K     52.00 K -0.00 % -0.00 % pv -L 25k
2019-08-02 03:11:21    50 be/4 root          0.00 K     80.00 K -0.00 % -0.00 % pv -L 25k
2019-08-02 03:11:22    50 be/4 root          0.00 K    104.00 K -0.00 % -0.00 % pv -L 25k
2019-08-02 03:11:23    50 be/4 root          0.00 K    128.00 K -0.00 % -0.00 % pv -L 25k
```

Toradex
Embedded. Computing.

blktrace/blkparse

- Overwhelming Amount of Output
- Make use of Filters
- Goal: Tracking Userspace PID Once Write to Flash is Confirmed
- C (Complete): Request Completed (Details Sector, Request Size and Success/Failure)
- I (Inserted): Request Sent to I/O Scheduler for Addition to Internal Queue

```
root@colibri-imx6:~# blktrace -o - /dev/mmcblk1 | blkparse -i -
179,0    0        26     0.000114661   304  A  WS 4509800 + 8 <- (179,2) 4468840
179,0    0        27     0.000117328   304  Q  WS 4509800 + 8 [jbd2/mmcblk1p2-]
179,0    0        28     0.000119661   304  M  WS 4509800 + 8 [jbd2/mmcblk1p2-]
179,0    0        29     0.000127328   304  U   N [jbd2/mmcblk1p2-] 1
179,0    0        30     0.000131661   304  I  WS 4509736 + 72 [jbd2/mmcblk1p2-]
179,0    0        31     0.008860277   279  D  WS 4509736 + 72 [kworker/0:3H]
179,0    0        32     0.012586780   279  C  WS 4509736 + 72 [0]
```

barrier	barrier attribute
complete	completed by driver
fs	FS requests
issue	issued to driver
pc	packet command events
queue	queue operations
read	read traces
requeue	requeue operations
sync	synchronous attribute
write	write traces
notify	notify trace messages

Toradex
Embedded. Computing.

Lifespan Estimation

- Logging Flash Health and I/O Tracking
- Storing in Local Database
- Correlations:
  - Flash Health Over Time

$$lifespan \text{ in } seconds = \frac{endurance}{average\ global\ block\ erase\ count}$$

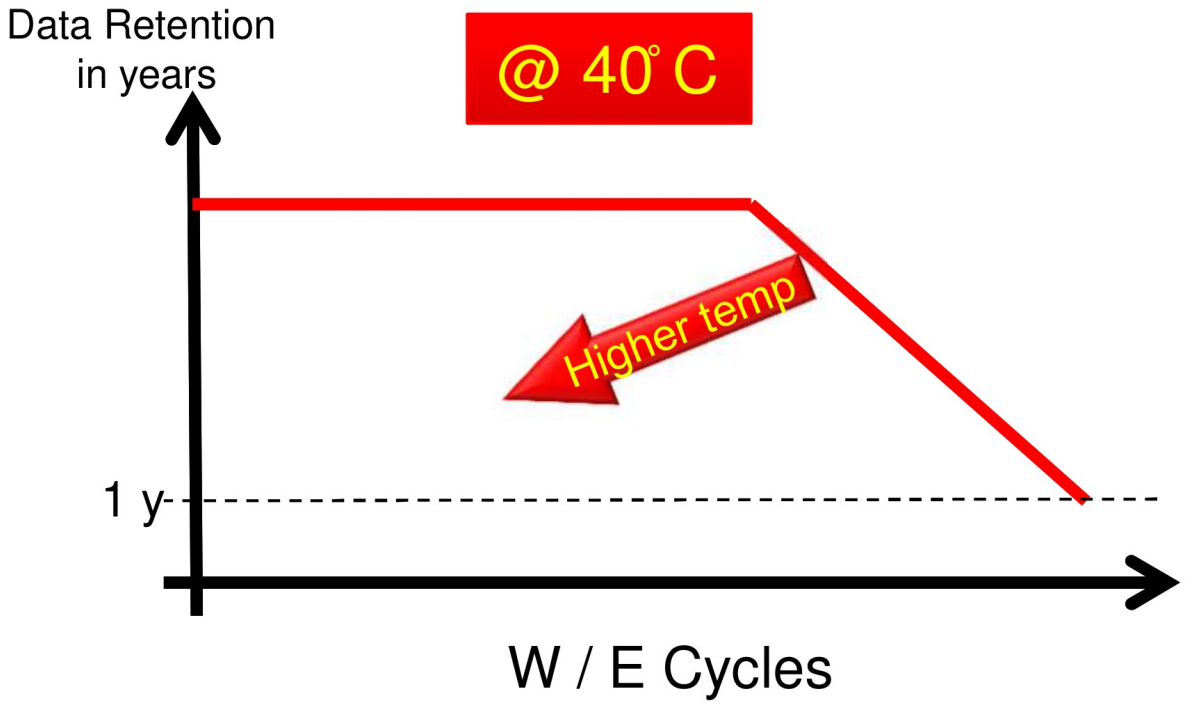
  - Flash Health Dependent on Write Rate

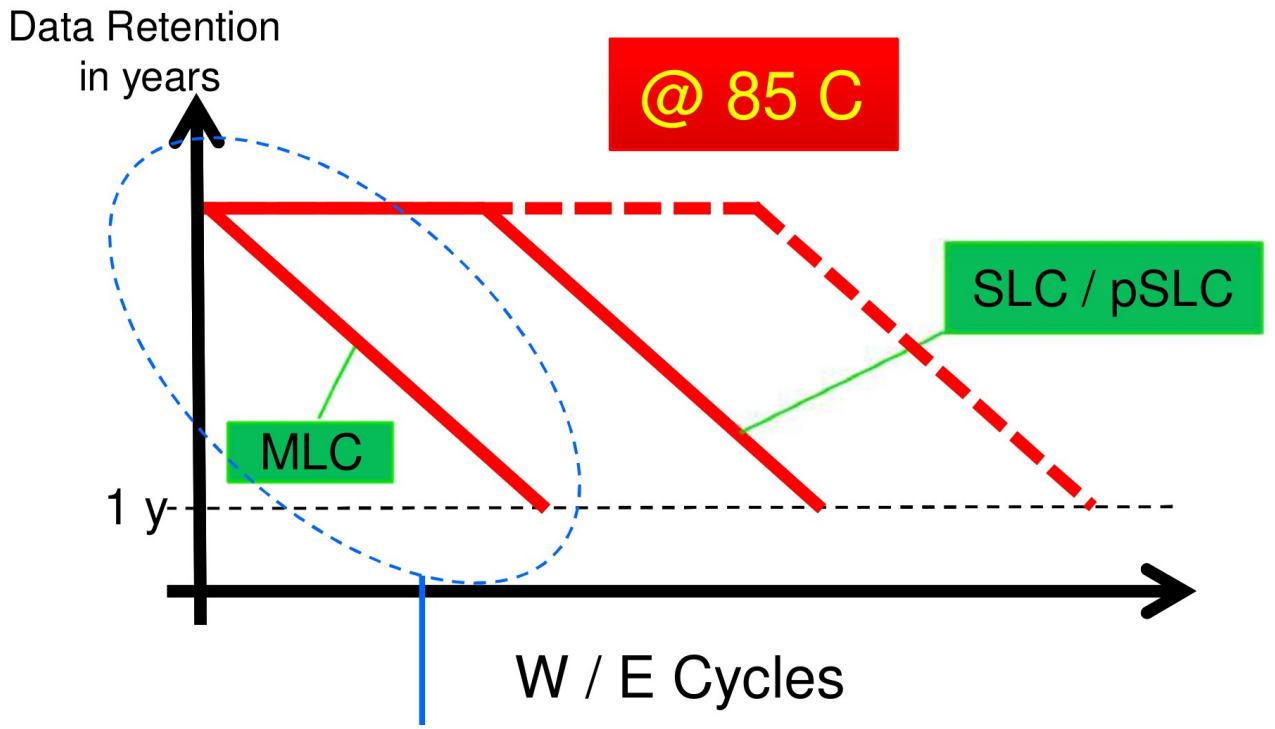
$$lifespan = \frac{endurance}{adjusted\ average\ write\ rate}$$

Toradex
Embedded. Computing.

Remark on Wear Estimation

- Temperature Strongly Affects Flash Lifespan!





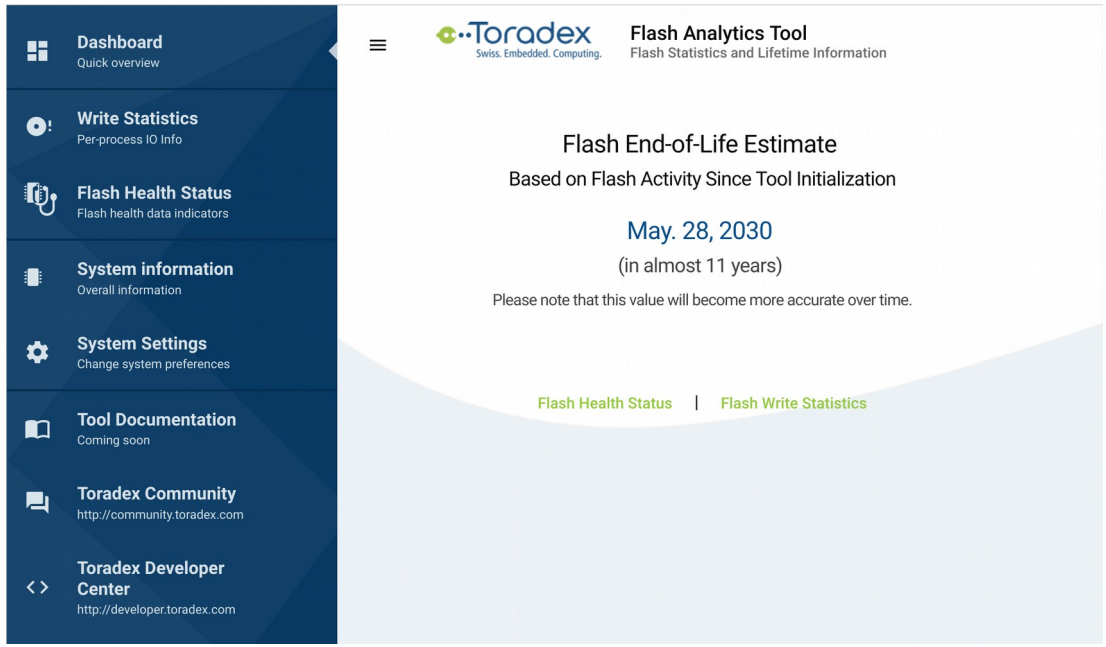
Toradex
Embedded. Computing.

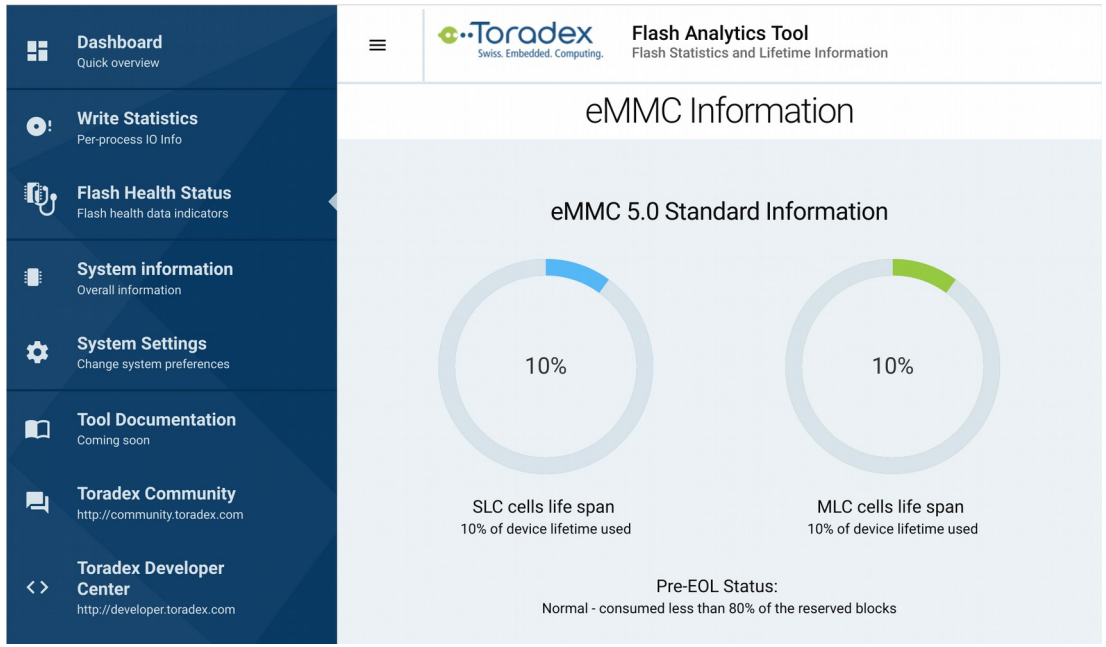
Flash Analytics Tool

- Under Development at Toradex Labs
- Abstracting Away Complexity of Wear Estimation
- Targeting Application Developers
- Current Prediction Model Implemented Using Linear Regression

Toradex
Embedded. Computing.

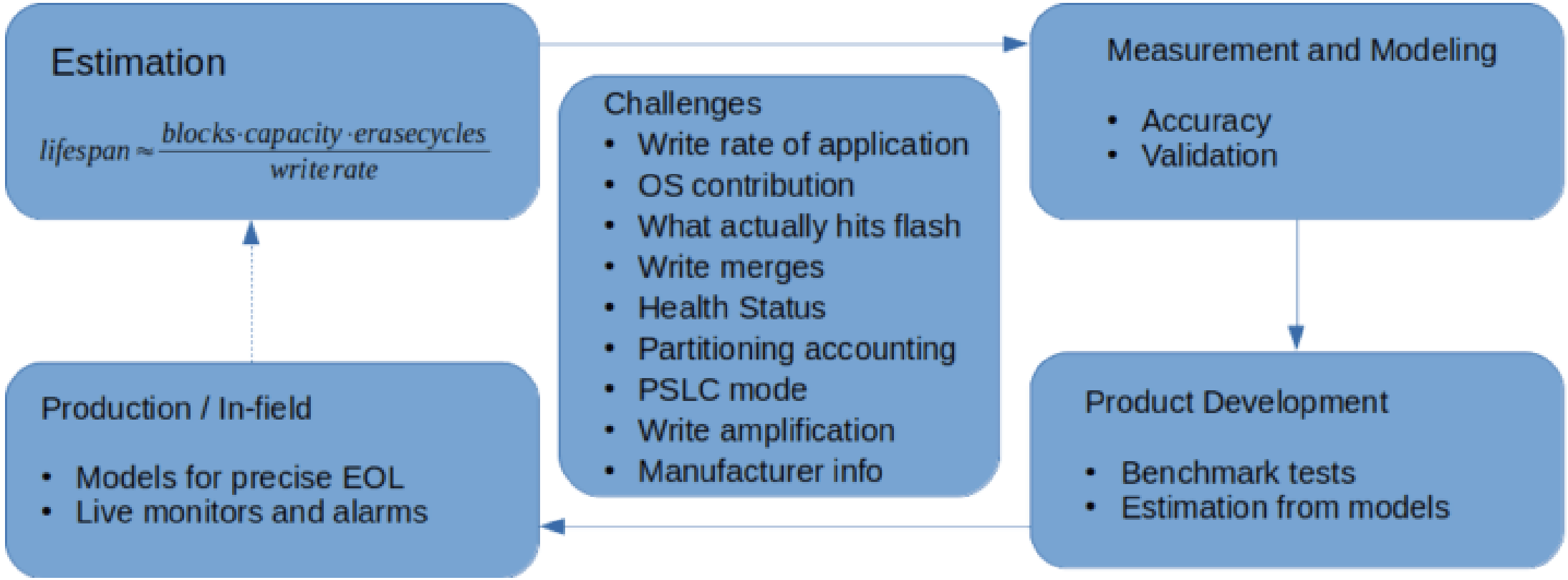
Live Demo





Toradex
Embedded. Computing.

Conclusion



Toradex
Embedded. Computing.

Questions ?

Toradex
Embedded. Computing.

References

- Toradex Labs - https://labs.toradex.com
- Flash Analytics Tool - https://labs.toradex.com/projects/flash-analytics-tool
- Toradex blog – What you should know about Flash storage - https://www.toradex.com/pt-br/blog/what-you-should-know-about-flash-storage
- Flash Memory - Wikipedia - https://en.wikipedia.org/wiki/Flash_memory
- Micron NOR | NAND Flash Guide
- Micron Choosing the Right NAND - https://www.micron.com/products/nand-flash/choosing-the-right-nand
- Flash 101: NAND Flash vs NOR Flash - https://www.embedded.com/design/prototyping-and-development/4460910/Flash-101--NAND-Flash-vs-NOR-Flash
- Cactus Technologies White Paper - CTWP016: An Overview of Pseudo-SLC NAND - https://www.cactus-tech.com/files/cactus-tech.com/documents/whitepapers/An%20Overview%20of%20Pseudo-SLC%20NAND.pdf
- Cactus Technologies SLC, pSLC, MLC and TLC Differences - Does Your Flash Storage SSD Make the Grade? - https://www.cactus-tech.com/resources/blog/details/slc-pslc-mlc-and-tlc-differences-does-your-flash-storage-ssd-make-the-grade
- 11 Myths About NAND Flash - https://www.electronicdesign.com/memory/11-myths-about-nand-flash
- How NAND flash degrades and what vendors do to increase SSD endurance - https://searchstorage.techtarget.com/podcast/How-NAND-flash-degrades-and-what-vendors-do-to-increase-SSD-endurance
- Micron TN-29-42 – Wear-Leveling Techniques in NAND Flash Devices
- Wear Leveling - Wikipedia - https://en.wikipedia.org/wiki/Wear_leveling
- Micron TN-2960: Garbage Collection in SLC NAND Flash Memory
- MultiMediaCard - Wikipedia - https://en.wikipedia.org/wiki/MultiMediaCard
- Embedded Multi-Media Card (e.MMC) Electrical Standard (5.0) - https://www.jedec.org/sites/default/files/docs/JESD84-B50.pdf
- Macronix Application Note Managing Unexpected NAND Flash Power Loss in Embedded Systems - http://www.macronix.com/Lists/ApplicationNote/Attachments/1924/AN0363V1%20-%20Managing%20Unexpected%20NAND%20Flash%20Power%20Loss%20In%20Embedded%20Systems.pdf
- Micron TN-FC-32: e.MMC Device Health Report
- Toshiba NAND Flash Memory Solutions - http://igexact.org/storage/legacy/uploads/files/FG_ENG/20170329/Toshiba%20NAND%20Flash%20Memory%20Solutions%20-%20Product%20Introduction%20-%20Exact%20Event%20March%202017.pdf

Toradex
Embedded. Computing.

References Continued

- The Linux IO Stack unveiled - Thomas Schöbel-Theuer - http://www.linuxtag.org/2013/fileadmin/www.linuxtag.org/slides/Thomas_Schoebel-Theuer_-_Der_Linux_I_O-Stack.e201.pdf
- Linux block I/O tracing - Gabriel Krisman Bertazi - https://www.collabora.com/news-and-blog/blog/2017/03/28/linux-block-io-tracing/
- Budget Fair Queueing (BFQ) Storage-I/O Scheduler - http://algo.ing.unimo.it/people/paolo/disk_sched/
- Deadline scheduler - Wikipedia - https://en.wikipedia.org/wiki/Deadline_scheduler
- Noop scheduler - Wikipedia - https://en.wikipedia.org/wiki/Noop_scheduler
- The Linux Kernel/Storage - Wikibooks - https://en.wikibooks.org/wiki/The_Linux_Kernel/Storage
- I/O Scheduling - Wikipedia - https://en.wikipedia.org/wiki/I/O_scheduling
- An Introduction to Linux Block I/O – Avishay Traeger - https://researcher.watson.ibm.com/researcher/files/il-AVISHAY/01-block_io-v1.3.pdf
- Understand your NAND and drive it within Linux – Miquèl Raynal - https://archive.fosdem.org/2018/schedule/event/nand_on_linux/attachments/slides/2576/export/events/attachments/nand_on_linux/slides/2576/raynal_drive_your_nand_within_linux.pdf
- MTD stack documentation - http://www.linux-mtd.infradead.org/doc/general.html
- UBI – Unsorted Block Images - http://www.dubeiko.com/development/FileSystems/UBI/ubidesign.pdf
- UBI headers - http://www.linux-mtd.infradead.org/doc/ubi.html#L_ubi_headers
- UBIFS FAQ and HOWTO - http://www.linux-mtd.infradead.org/faq/ubifs.html
- UBIFS – UBI File-System - http://www.linux-mtd.infradead.org/doc/ubifs.html
- Linux Page Cache Basics - https://www.thomas-krenn.com/en/wiki/Linux_Page_Cache_Basics
- Don’t fear the fsync - http://thunk.org/tytso/blog/2009/03/15/dont-fear-the-fsync/
- The future of the page cache - https://lwn.net/Articles/712467/
- Micron TN-FC-25: Understanding Linux Driver Support for e.MMC
- Linux Tracing Technologies - https://www.kernel.org/doc/html/latest/trace/index.html#
- Using the Linux Kernel Tracepoints - https://www.kernel.org/doc/html/latest/trace/tracepoints.html
- Block I/O Layer Tracing: blktrace - Alan D. Brunelle - https://www.mimuw.edu.pl/~lichota/09-10/Optymalizacja-open-source/Materialy/10%20-%20Dysk/gelato_ICE06apr_blktrace_brunelle_hp.pdf
- blktrace User Guide - Alan D. Brunelle - http://www.fis.unipr.it/doc/blktrace-1.0.1/blktrace.pdf

Toradex
Embedded. Computing.

THANK YOU FOR YOUR INTEREST.

www.**toradex**.com | **developer**.toradex.com | **community**.toradex.com | **labs**.toradex.com