Microsemi
Power Matters.

Prioritization in the Linux NVMe Block Layer

Jon Saunders - Microsemi Corporation

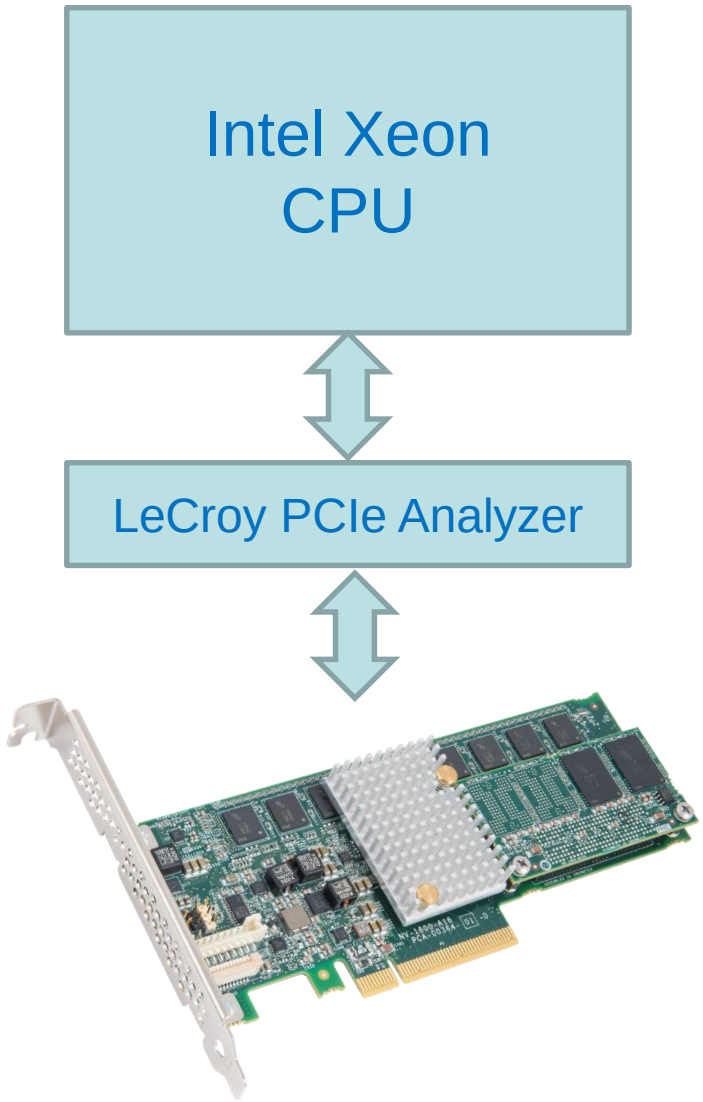
Stephen Bates – Microsemi Corporation

Background

- Improved NVMe block layer performance has become critical with the increasing NVMe storage speeds and decreases in latency.
- The Linux community is working to address this need by adding polling to the NVMe block layer, to create a way to prioritize NVMe Block IO by polling for completions.

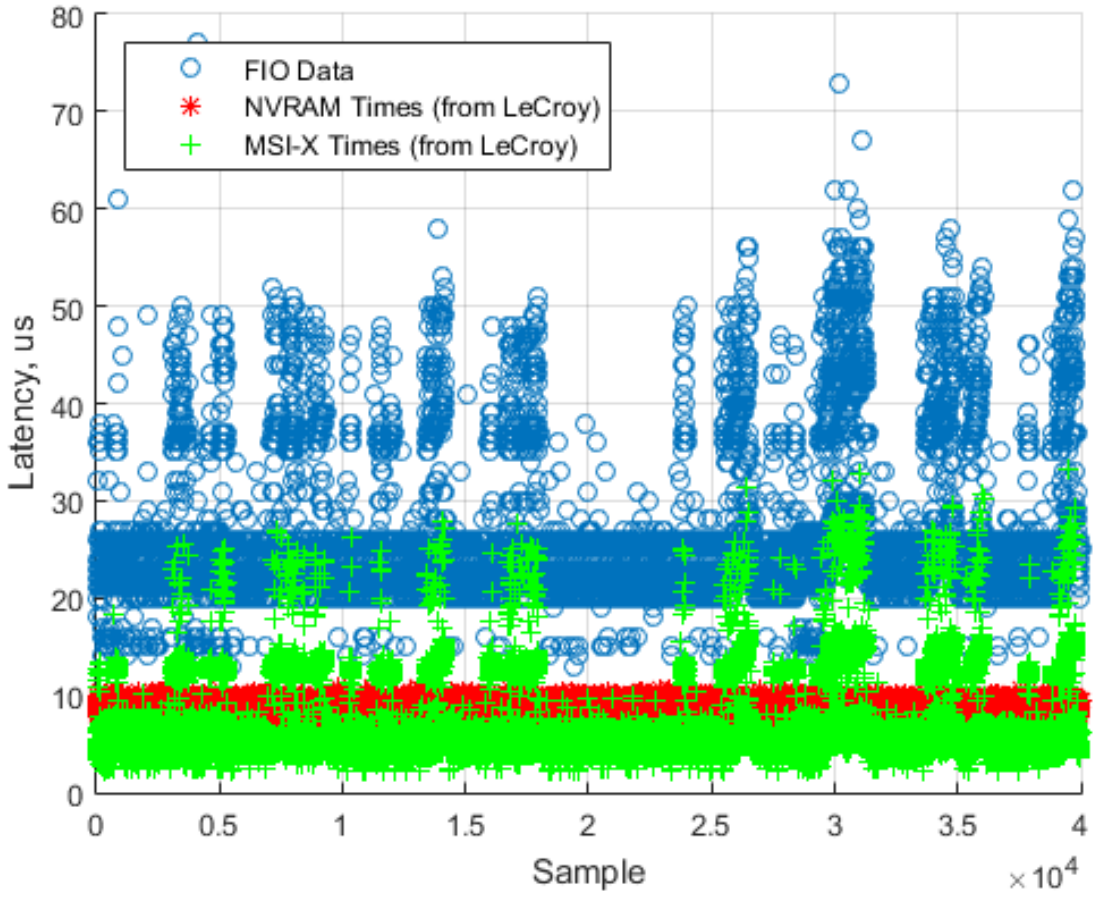
Latency

- When latency goes below 20 µs, it is important to analyze the factors contributing to that latency
  - Storage
  - OS
  - CPU architecture



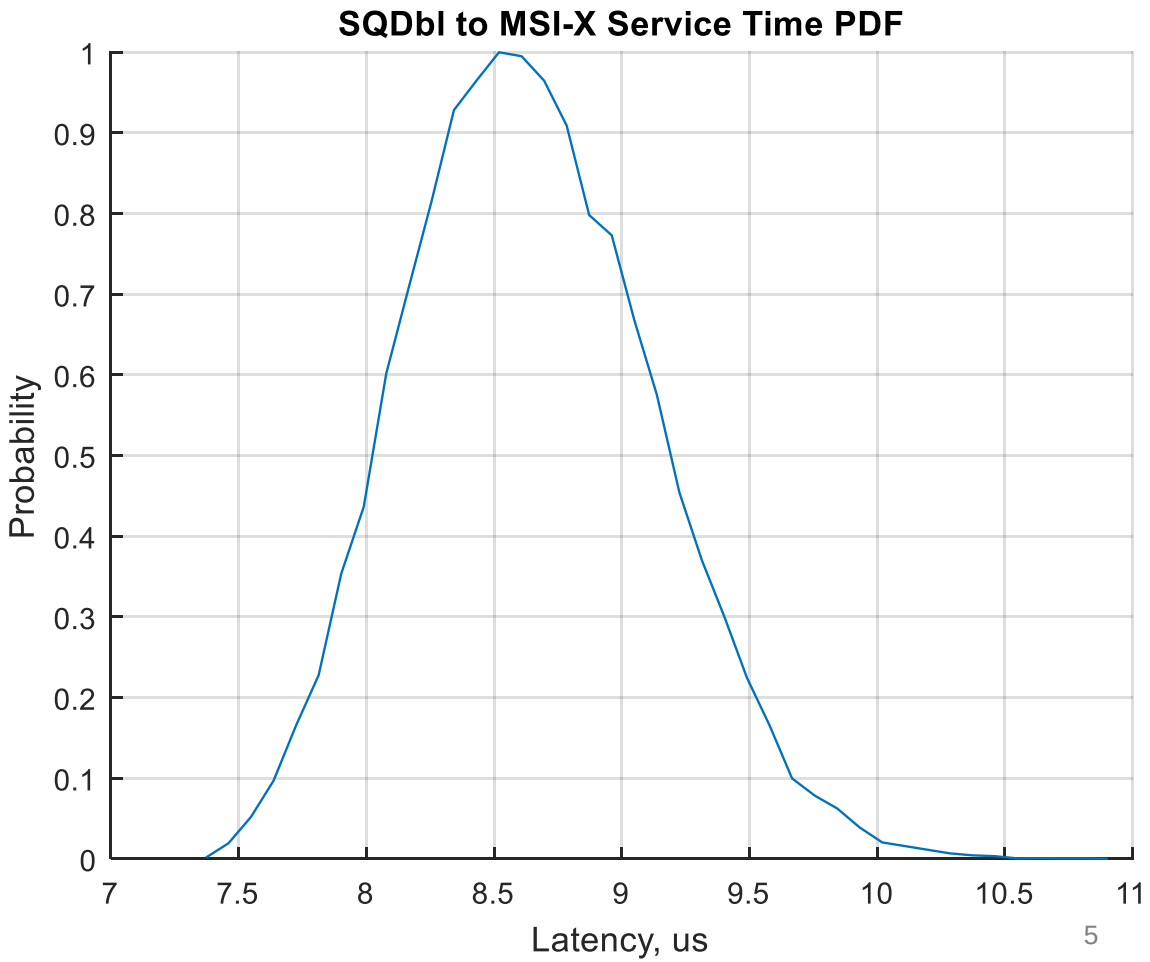
Latency Analysis

- The latency contribution from the NVRAM is consistently between 8 µs - 9 µs
- The latency measured at the application is ~10 µs more (this is mostly due to software overhead).
- The application overhead is also more “spikey”.



NVRAM Latency Analysis

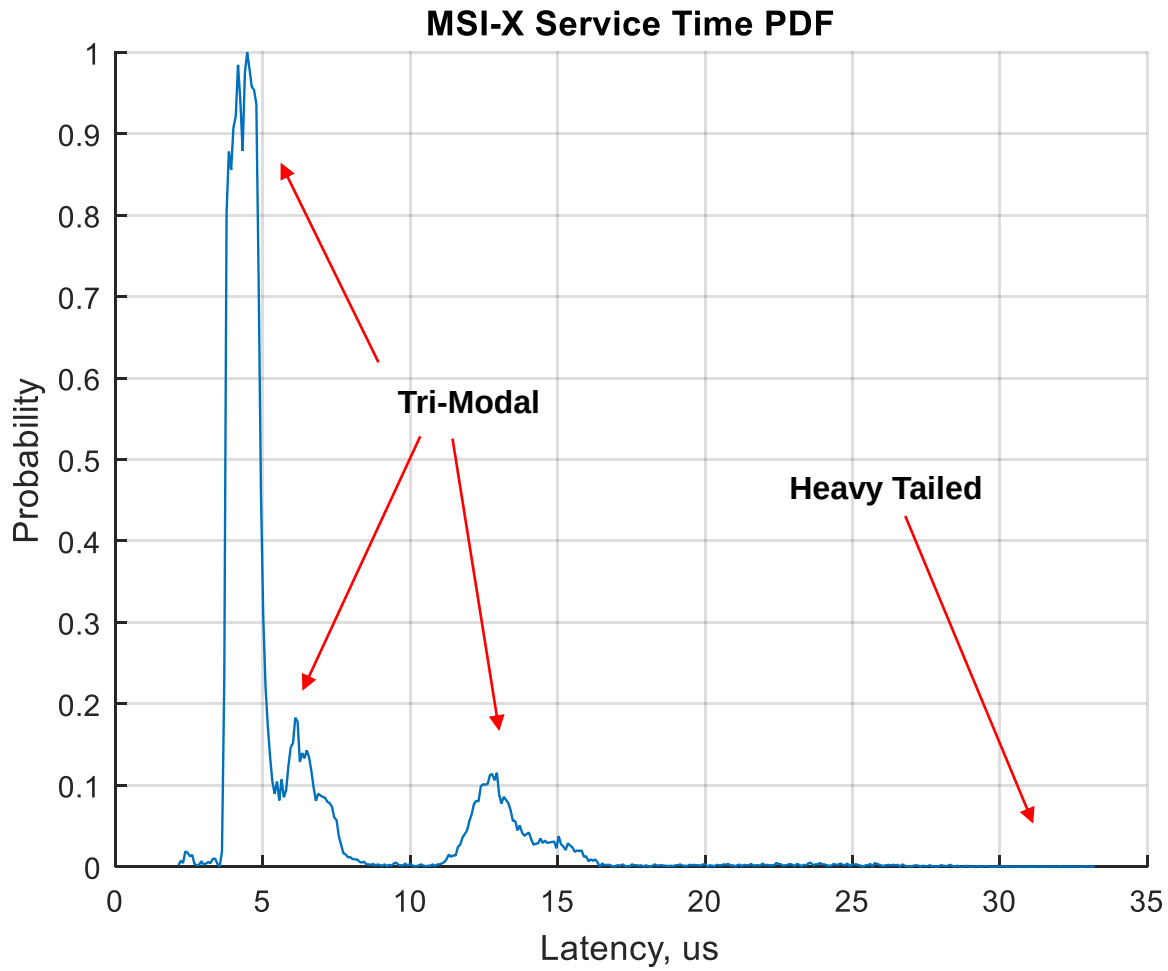
The following diagram shows Microsemi Flashtec™ NVRAM latency as a distribution. The distribution is very tight and consistent, with a mean latency of under 9 µs



MSI-X Latency Analysis

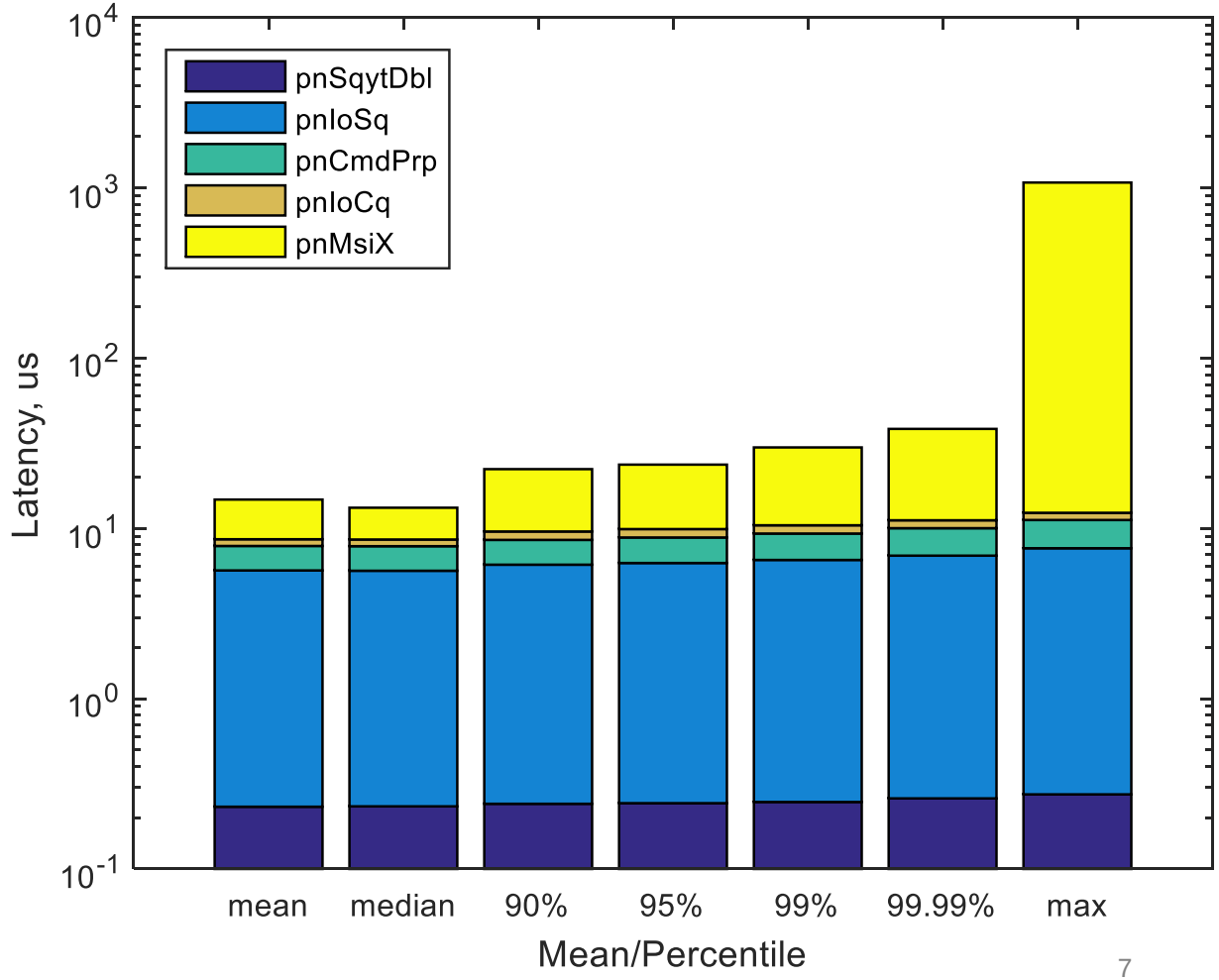
The following diagram shows the MSI-X service time as a distribution.

- The MSI-X service time is a function of the CPU architecture and OS (that is, it has nothing to do with the storage).
- The distribution is very loose, with a heavy tail as well as outliers.



Latency with MSI-X: The Need for Polling

- MSI-X servicing can be problematic when working with very low latency storage (when access times drop to a few tens of µs or less), severely impacting QoS by introducing outliers to the latency distribution.
- Introducing polling removes the need for MSI-X interrupts.

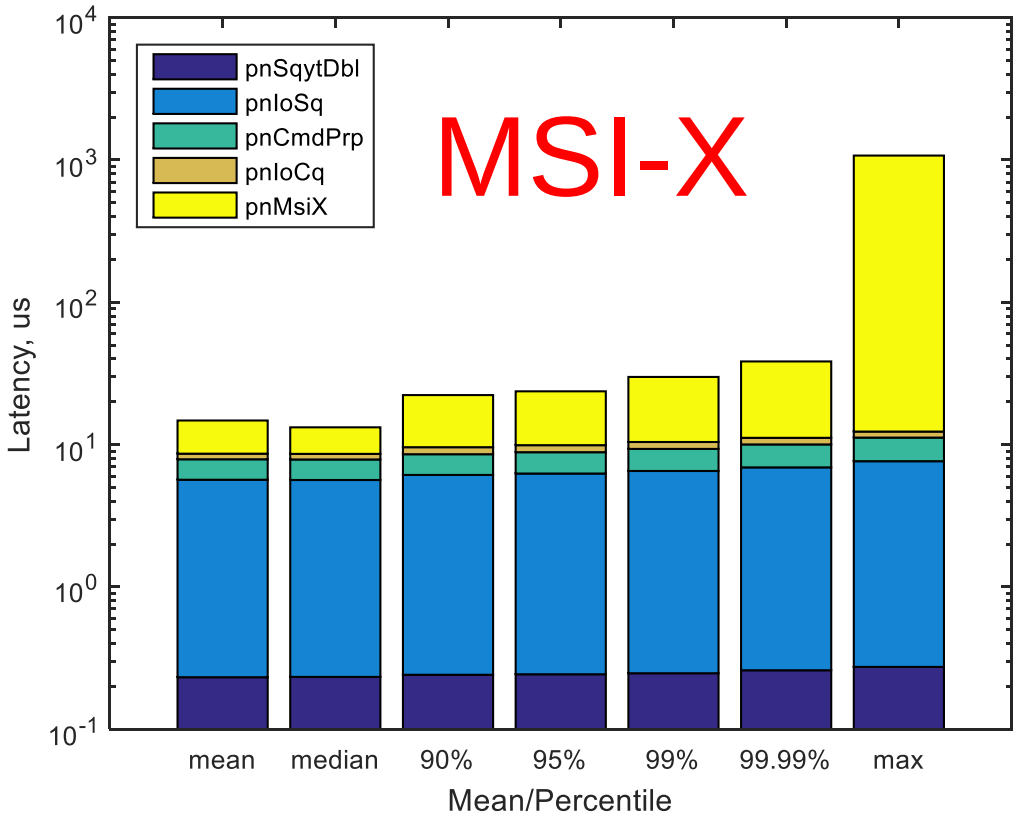


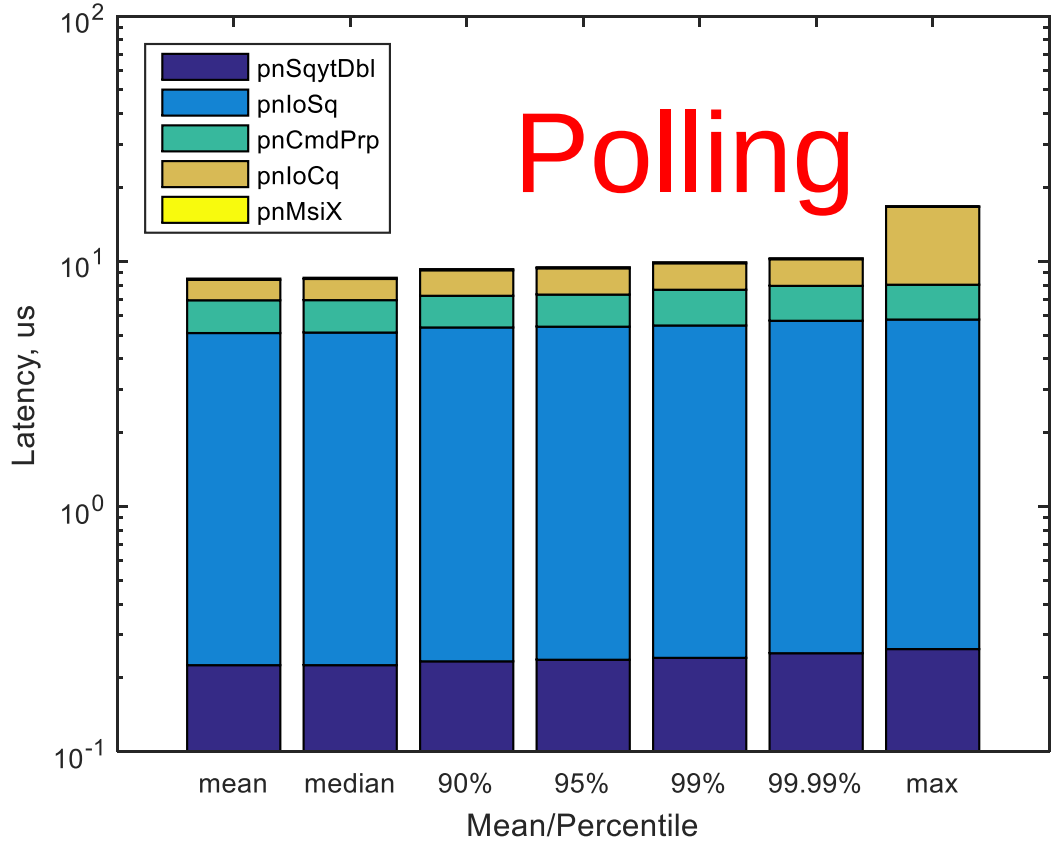
Polling in Linux

- Polling was added to the Linux 4.5, but only for NVM Express devices. The initial version of polling was global: either ON or OFF per H/W queue.

- We have experimented with this commit to look at the benefits of polling across the entire SSD.

Benefits of Polling





Polling reduces average latency to a limited degree.

Polling reduces the outlier latency by a large degree.

How it Will Work in Linux

- Applications will be able to flag individual IO as either normal priority (interrupt-based) or high priority (polling-based)
- As of Linux 4.6, IO will be passed into the kernel using preadv2() and pwritev2() system calls.
- When high priority IO are on a H/W queue, then that queue will be polled for completion (thus other IO may benefit).
- Still some work to be done to tie this all together and tune for performance

Conclusions

- NVMe storage speeds continue to increase, and new memory technologies such as 3DX and ReRAM will only accelerate this.
- NVMe storage can service random read and write IO consistently in under 20 µs. At this point, the overhead of MSI-X interrupts becomes very noticeable.
- Linux is adding support for polling for completions to reduce latency and latency outliers.
- Polling will be possible on a per-IO basis, but there is still some work to be done to deliver the capability to applications.

Flash Memory SUMMIT

Microsemi
Power Matters.

THANKS!!

Microsemi

Visit us at booth #213

www.microsemi.com