Slides: LA-UR-21-24529 Video: LA-UR-21-24530

More Performant Cluster State Management Using Open Source Firmware and a Kraken





Devon T. Bautista, J. Lowell Wofford

Los Alamos National Laboratory



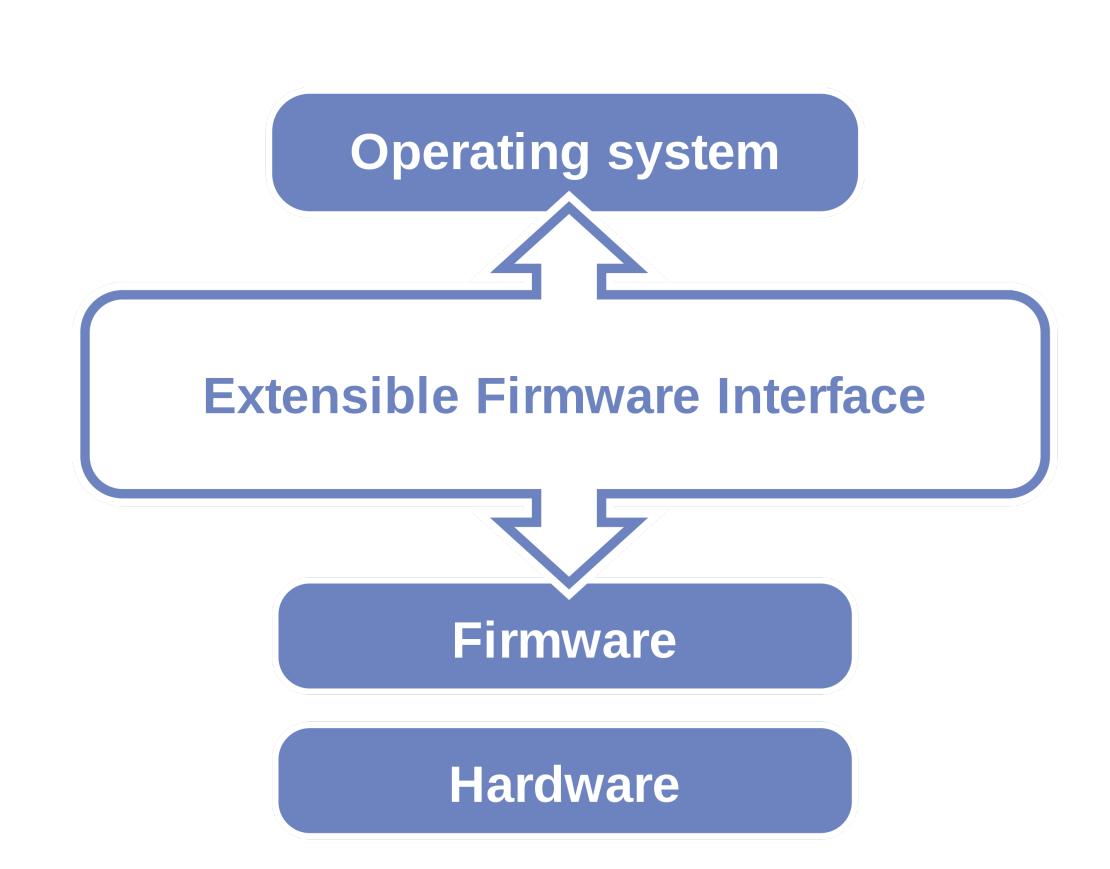
01 June 2021

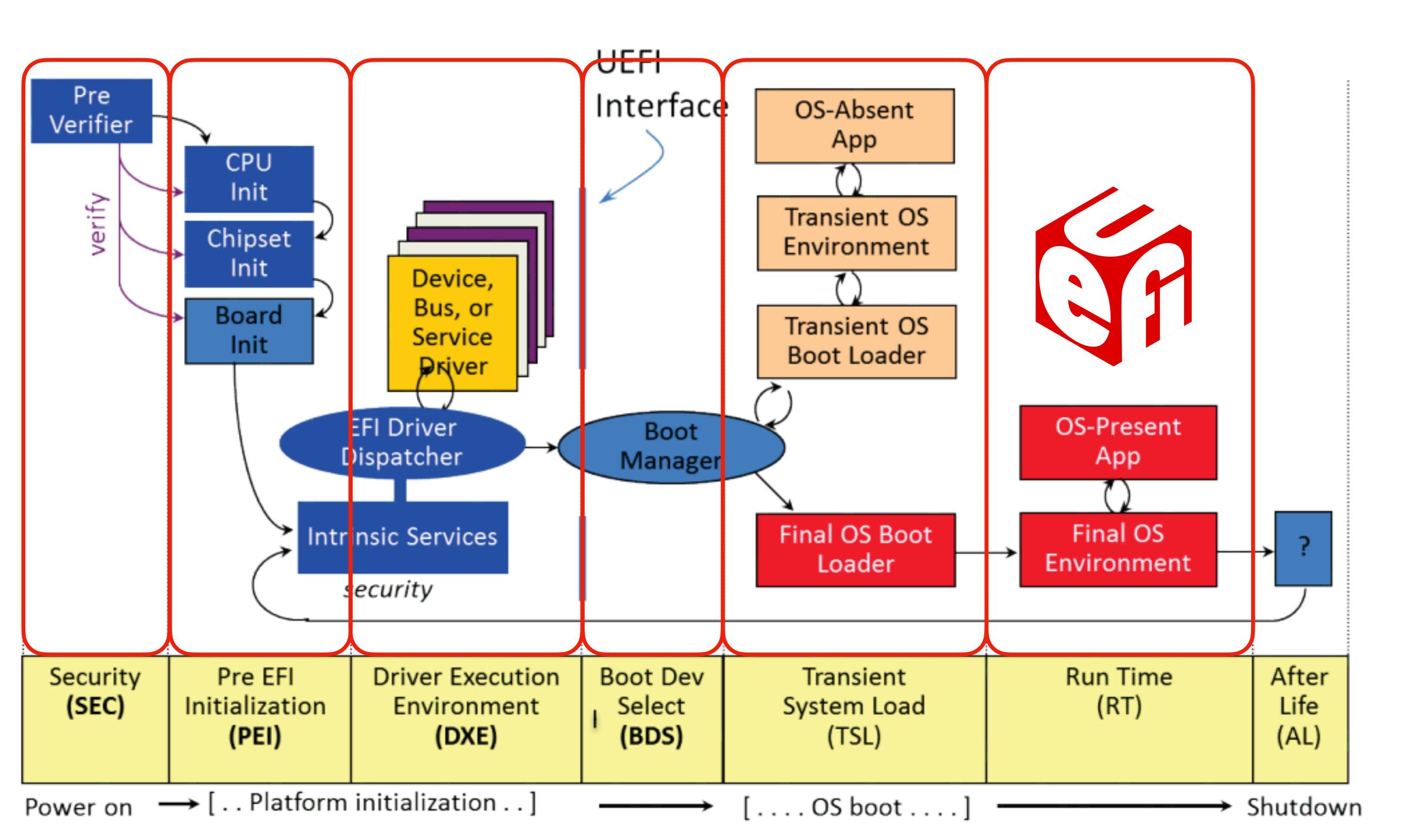
A Little Background on Firmware



UEFI: How We Got Here BIOS mechanism:

- Mundane
 - Control given to first boot loader found
 - Blindly executes
- Proprietary
 - Sans Libreboot, Coreboot
- Limited
 - 16-bit real-mode addressing
 - Operates in up to 1 MB of space
 - Programmed in assembly language
 - Can only address up to 2.2 TB drives
- 1998: "Extensible Firmware Interface"
- 2005: *"Unified* Extensible Firmware Interface"





Towards Open Source Firmware

DxeCore FastVideoDxe cdDxe RegAccessDxe egAccessSMM eportStatusCodeRouterRuntimeDxe tatusCodeHandlerRuntimeDxe leportStatusCodeRouterSmm StatusCodeHandlerSmm)atahubStatusCodeHandlerDxe StatusCodeRuntimeDxe

SectionExtractionD: enericIpmi SmmGenericIpmi baConfigDatabaseD baInitDxe

StaticSkuDataDxeNeonCityEPRP SetupConfigUpdateDxeNeonCityEPRP promUpdateDxeNeonCityEPRP SmbiosDataUpdateDxeNeonCityEPRP sbOcUpdateDxeNeonCityEPRP ioCfgUpdateDxeNeonCityEPRP SlotDataUpdateDxeNeonCityEPRP sidSvidDataUpdateDxeNeonCityEPRP StaticSkuDataDxeOpalCitySTHI SetupConfigUpdateDxeOpalCitySTHI promUpdateDxeOpalCity5TH SmbiosDataUpdateDxeOpalCityS711 sb0cUpdateDxe0palCityS ioCfgUpdateDxeOpalcitySTHI SlotDataUpdal.oxeOpalCitySTHI sidSvidPataUpdateDxeOpalCitySTHI StaticSkuDataDxeWolfPass SetupConfigUpdateDxeWolfPass promUpdateDxeWolfPass SmbiosDataUpdateDxeWolfPass sbOcUpdateDxeWolfPass lioCfgUpdateDxeWolfPass lotDataUpdateDxeWolfPass chUsb2EyeDiagramDxeWolfPass SsidSvidDataUpdateDxeWolfPass pkConfigUpdateWolfPac

DMIMarginUpdateDxeW SwitchLedWA StaticSkuDataDxeBu

SetupConfigUpdateD promUpdateDxeBuchananPass SmbiosDataUpdateDxeBuchananPass sb0cUpdateDxeBuchananPass ioCfgUpdateDxeBuchananPass lotDataUpdateDxeBuchananPass iserPCIeBifurcationDxeBuchananPass chUsb2EveDiagramDxeBuchananPass SsidSvidDataUpdateDxeBuchananPass pkConfigUpdateBuchananPass BdsUpdateHookBuchananPass StaticSkuDataDxeLightningRidgeEXRP SetupConfigUpdateDxeLightningRidgeEXRP OpromUpdateDxeLightningRidgeEXRP SmbiosDataUpdateDxeLightningRidgeEXRP Usb0cUpdateDxeLightningRidgeEXRP IioCfgUpdateDxeLightningRidgeEXRP

SlotDataUpdateDxeLightningRidgeEXRP SsidSvidDataUpdateDxeLightningRidgeEXRP_Cpulo2Dxe StaticSkuDataDxeSawtoothPass SetupConfigUpdateDxeSawtootPass OpromUpdateDxeSawtoothPass SmbiosDataUpdateDxeSawtoothPass Usb0cUpdateDxeSawtoothPass IioCfgUpdateDxeSawtoothPass SlotDataUpdateDxeSawtoothPass SsidSvidDataUpdateDxeSawtoothPass **FpkConfigUpdateSawtoothPass**

Uefi0promSetup

IioCfgUpdateDxeNeonCityFPGA SIotDataUpdateDxeNeonCityFPGA StaticSkuDataDxeBlueMountainPass figllodateDxeBlueMou

OemBadgingDxe JpegDecoderDxe ItkLogoProcess

LIOASDWW UuidDxe SpdRawData SyncSetupVariableToPcd PlatformStatusCodeHandlerDxe PlatformStatusCodeHandlerSmm SysconfigSysinfoHide PlatformVariableInitDxe CommonErrorGeneration ErrorManagerSetup PostErrorToSel

1sft0emActivation

lanagerMenuApp Boo. Ipm ootOrder tupMarkerFileBootOrder aticBootOrder tworkBootApp rontPanelLockout SataSgpio SmbiosPcTable SecureBootErrorHandlerUXe SecureBootProvisionDxe SecureBootSetup SecureBootCtrlDxe BiosGuardServices MeSmmProtocolThunk S3NvramSave PlatformEarlyDxe

CpuIoDxe HiiDatabase DataHubDxe FrameworkHiiAlias Legacy8259 CpuArchDxe PlatformCpuPolicy uMpDxe SMB SFilter Metrone Watchdog PcRtc RuntimeDxe PciHostBridge PpmInitialize IioSmm LpcPlatform PilotIVPc8374

ReserveMen kService ckServiceSmm /InterruptHook[BiosReverseThu jacy acyInstallInt1 ultTolerantWrit bleSmm iableInt15 riableInstallIn

MonotonicCounterRun BdsDxe SecurityStubDxe JpegDecoderDxe OemBadgingDxe CapsuleRuntimeDxe RandomPoweronRacks AcmErrorReport Msft0emActivation TpmPlatformDxe CpPcBiosId SotupBmcCfg

BootTimeOut derPolicyDxe iButton NmiButton Smbios0obMdr1 SmbiosOobMdr2 MemoryMappingOob OemPostScreenDisplay SolStatusPlatform ConsoleBdsUpdate DualVideo PayloadBoot UefiNetworkStackProtocolDxe UetiOpromSetup Enter BootManagerSetup BootMaintSetup DriverHealthDxe FdUpdate

VlanConfigDxe ArpDxe Dhcp4Dxe Ip4Dxe Mtftp4Dxe Udp4Dxe Tcp4Dxe UefiPxeBcDxe IScsiDxe Ip6Dxe TcpDxe Udp6Dxe Dhcp6Dxe Mtftp6Dxe HttpDxe HttpBootDxe HttpUtilitiesDxe DnsDxe

> PchSpiSmm PchSerialGpio Smartlimer PchResetRuntime WdtDxe PlatformReset PowerButtonHandler TegMor Tcg2Dxc Tcg2Smm



Tr dDxe Smm AcpiSwChild tformPreVariableDxe Dxe Elog tformErrorGeneration

ericElog BmcElog GenericElog iRecirFru VariableDriver MeFwPatch Status Driver nericFru BusDxe AtapiPassThru elRaidAtaAtap elRaidBiosThu ompatiblePciD BusDxe SerialDxe FloppyDxe acyBiosDxe CVideo cfgSyncDxe Platform acyBiosPlatfor olicy DIMMDriver DIMMHii AAHCI eSATALegacy esSATALegacy eSATAEFI esSATAEFI VR0C2 VR0C1 PcieGen3 CVideoGop VBIOS eedVideo titionDxe BusDxe orySubClass upBrowser Setup tformDevsUpdate

iosDxe

ScsiBus

ScsiDisk

SmmAccess

PiSmmIpl

PiSmmCore SmmCommunicationBuffer PiSmmCpuDxeSmm CpuIo2Smm BIOSGuard Ps2KeyboardDxe Ps2MouseDxe

WheaElog UncoreErrorLog PcieErrorLog PvMode KtiErrLogRuntime KtiErrorLogPost WheaPlatformBoot

IsaBusDxe IsaFloppyDxe IsaSerialDxe LegacyRegion2 LegacyBiosPlatform Legacy8259 LegacyInterrupt LegacyInterruptHookDxe LegacyBiosDxe

playEngine entDxe SmbiosMeasure EnglishDxe NvmExpressFxe ESRTIISata Eti AcpiPlatform

эхэшшнаноге CpuHotAdd KtiRas CpuRas **IioRas** HpIOXAccess RasInit PolicySampleDriver ImcErrorHandler ProcessorErrorHandler PcieErrorHandler PartialMirrorHandler EnhancedMcaErrorLog WheaErrorLogListener PprviseriorLogListener LegacyBridgeDxe WheaErrorInj2 RASMiscDriver McBankErrorIn QuiesceSupport

IsPlatfor apportunea WheeSapport AneaErrorInj

OPAPlatConfigSkt0 OPAP1atConfigSkt1 OPAPlatConfig_IFP45b_H79275 OPAPlatConfig IFT45b H79267 H79272 LsiVtdSupport.inf LegacyFV/CSM_v74 UsbRt Uhcd KbcEmulSMM KbcEmulDxe UsbInt13 AmiLegacyBiosHook LegacySredir

Y2KRollover

SmiGraphicsOutput



The $2\frac{1}{2}$ "Hidden OSes" on x86

Code you	R	ing 3 (User)		
know about				
	Ring 0 (Linux)			
	Ring -1 (Xen etc.)			
Code	Ring -2 kernel and ½ kernel Control all CPU resources. Invisible to Ring -1, 0, 3			
you				
don't know		SMM 1/2 kernel. Traps t 8086 16-bit mode.		
about		UEFI kernel running in 64-bit paged mode.		

X86 CPU you know about

Ring -3 kernels

Management Engine, ISH, IE. Higher privilege than Ring -2. Can turn on node and reimage disks invisibly. Minix 3.

X86 CPU(s) you don't know about

Redundant Drivers

	EDK2 Firmware	GI
Filesystem	↓ tianocore / edk2 <> Code 11 Pull requests 5	git
	양 master	summary path: root/g
	tianocore / edk2	
USB	<> Code 11 Pull requests 5 Actions Projects	Security 🗠 Insi
	양 master → edk2 / MdeModulePkg / Bus / Usb / Usb	BusDxe / UsbBus.c
Network	↓ tianocore / edk2 <> Code 11 Pull requests 5 Actions Proje	git
	우 master → edk2 / <u>NetworkPkg</u> / TcpDxe /	summary path: root/

RUB Bootloader

index : grub.git

GNU GRUB

/	refs	log	tree	commit	diff	
/grub-core/fs/fat.c						

	cit	index : grub.git				
nsigl	9.0	GNU	GRUB			
	summary	refs	log	tree	commit	diff
.c	path: root/gr	ub-core	e/bus/us	sb/usb.c		



Linux Kernel

/ fs / fat / fat.h

1 2	/* SPDX-License-Identifier: GPL-2.0 */ #ifndef _FAT_H
3 4	#define FAT_H
5	<pre>#include <linux buffer_head.h=""> #include <linux nls.h=""></linux></linux></pre>

/ drivers / usb / core / usb.c

1 2	// SPDX-License-Identifier: GPL-2.0 /*
3	<pre>* drivers/usb/core/usb.c</pre>
4	*
5	* (C) Copyright Linus Torvalds 1999
6	* (C) Copyright Johannes Erdfelt 1999-2001
7	1. (C) Convertable Andress Col 1000

/ net	/ ipv4	/ tcp.c	
1	//	SPDX-License	-Identifier: GPL-2.0-or-later
2 3 4	*	INET	An implementation of the TCF
5	*		operating system. INET is i interface as the means of cc
6 7	*		Implementation of the Transm

Problems

- Redundant drivers with little-to-no code sharing
 - Increased attack surface
 - Loading the same drivers multiple times is slow
- The code with the highest privilege and control over hardware is audited the least

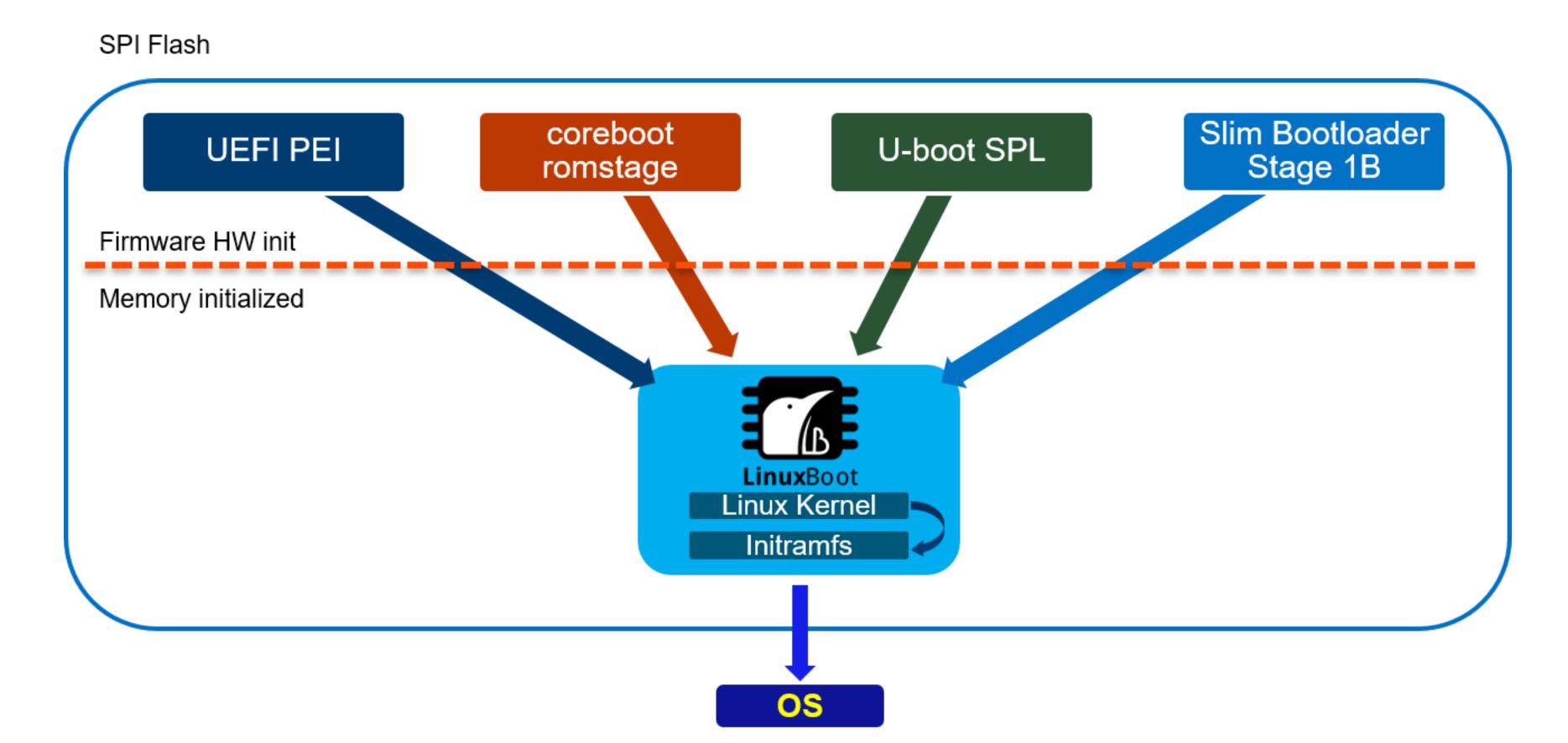
 - Proprietary, closed-source
- Reliance on vendor for updates and fixes
 - A bottleneck to your production timeline

• Less frequent update/deployment lifecycle than most software and operating systems

• Outsourcing development to *another* middleman in mitigating firmware updates/issues

Scary! What is being done?

"Let Linux do it." Replacing redundant, closed drivers with vetted, open ones



https://www.linuxboot.org

Open Firmware Using Linux

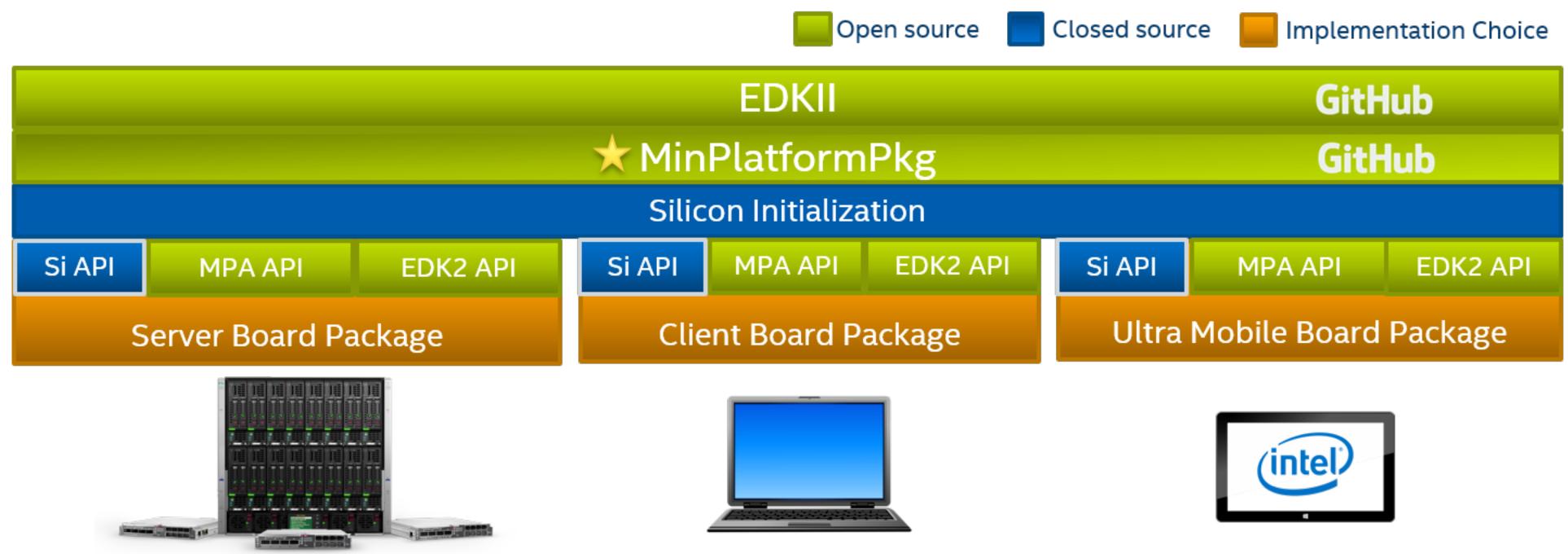
- One implementation of drivers
- Linux: Vetted for more than 20 years in military, consumer, and supercomputing systems
 - Already running on mission-critical devices around the world
 - Replace lightly-tested, closed drivers with hardened, heavily-tested, open source ones
- Bootstrap customization and fine-tuning for site-specific needs
 - More on the relevance to HPC upcoming
- More people understand Linux
 - Leverage existing talent/experience
- More mature tooling

Open Firmware at Facebook

- **2011:** Open Compute Project announced
- 2014: OpenBMC: Open source baseboard management controller firmware
 - Now a Linux Foundation project
- LISA18^[4]/OSFC 2018^[5]: Facebook uses Linuxboot in the cloud
 - "Booting is hard"
 - Many different types of devices now vs. one *de facto* standard then
 - More demands for firmware security"
 - Measured bootstrapping
 - "Provisioning is hard"
 - Firmware is now more complex
 - Need a robust provisioning solution

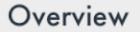
Intel





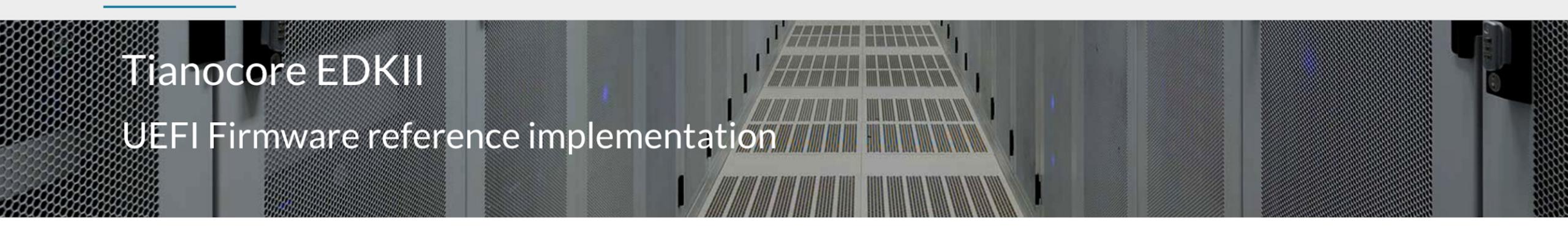
Source: https://software.intel.com/content/www/us/en/develop/articles/minimum-platform-architecture-open-source-uefi-firmware-for-intel-based-platforms.html





Building EDKII Firmware

Dynamic tables framework



Overview

Arm is an active contributor to the EDKII project hosted by the Tianocore community.

The EDKII project is an open source project that provides a modern, feature-rich, cross-platform firmware development environment for the UEFI and PI specifications developed and maintained by the UEFI Forum.

Arm contributions make sure the EDKII project constantly keeps an up to date implementation of a UEFI compliant firmware on Arm systems.

Arm contributes to both the EDKII main repository, maintaining some core packages like DynamicTablesPkg and StandaloneMMPkg, and the EDKII platforms repository, hosting support for various Arm reference platforms as well as other 3rd party Arm-based platforms maintained by either Linaro or partners.





Open Source Firmware in HPC

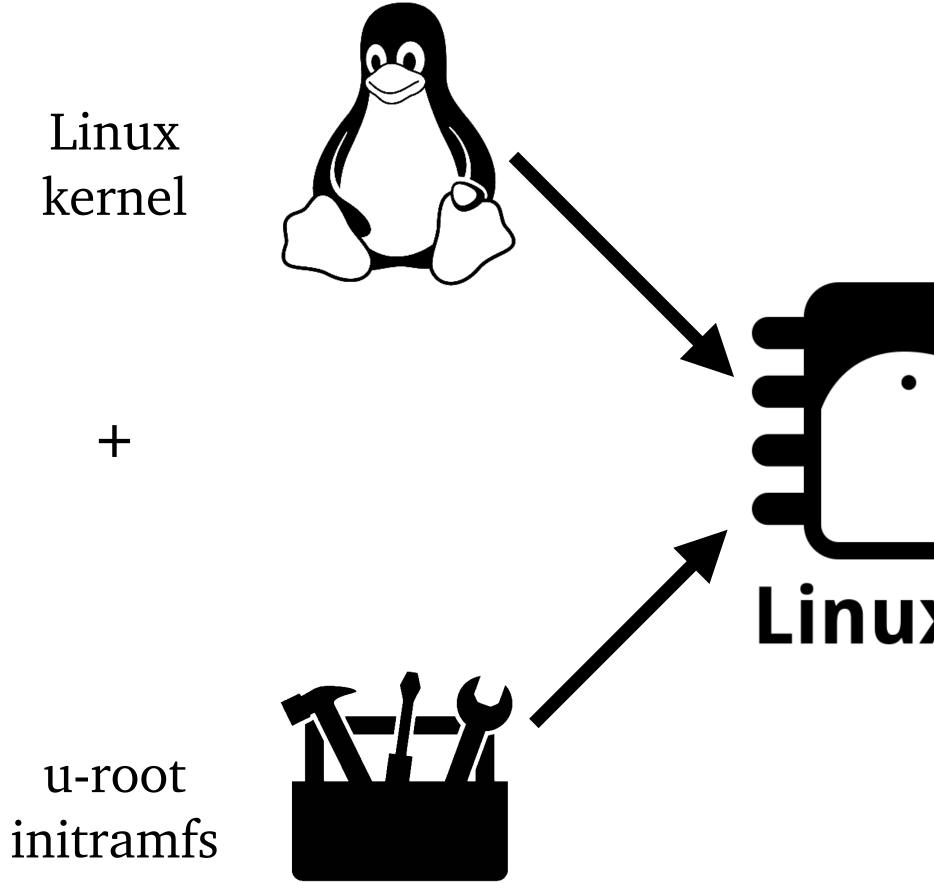
Motivation

- The need to boot systems most efficiently
 - Custom system initialization: greater control, finer performance
 - Boot times matter due to nodes more frequently rebooting
 - Vendor firmware is generic
- The need to use modern, secure protocols
 - TFTP, DHCP implementations can be buggy
 - ► HTTPS
 - TLS client certificates for cryptographic root-of-trust between nodes and parent
- The desire to run an extremely minimal operating system on compute nodes
 - Containerize user jobs directly on top of hardware
- the nodes

• The desire to potentially run a cluster state manager at a very low level to have better control of

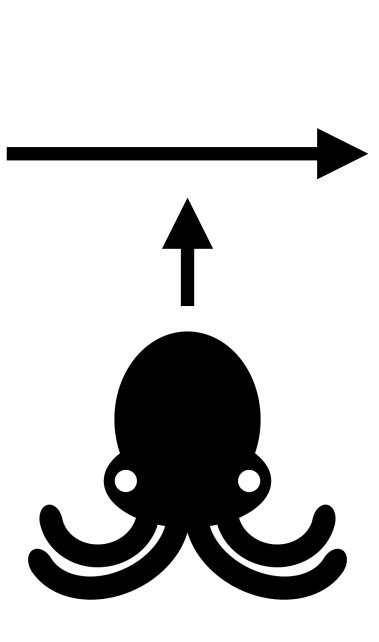


Open Firmware at LANL





LinuxBoot



kraken



- <u>http://kraken-hpc.io/</u>
- <u>https://github.com/kraken-hpc/</u>

Sources and Further Reading

- [1] Hudson, Trammell. "Bringing Linux back to the server BIOS with LinuxBoot". <u>https://trmm.net/LinuxBoot_34c3/</u>. 2017-12-29.
- [2] Ververis, Vassilios. "Security Evaluation of Intel's ActiveManagement Technology" <u>cover.pdf</u>.
- [3] Minnich, Ron. "Replace Your Exploit-Ridden Firmware with Linux". <u>https://youtu.be/iffTJ1vPCSo</u>. 2017-10-27.
- Reliable with LinuxBoot" *LISA18*. https://www.usenix.org/conference/lisa18/presentation/barberio. 2018-10-31.
- [5] Hendricks, David; Barberio, Andrea. "Open Source Firmware @ Facebook". OSFC 2018. <u>https://www.youtube.com/watch?v=eKVSBESoKUc</u>.
- *OSFC 2019. <u>https://www.youtube.com/watch?v=x3NFbUC3hkA</u>. 2019-12-02.*

https://people.kth.se/~maguire/DEGREE-PROJECT-REPORTS/100402-Vassilios Ververis-with-

[4] Hendricks, David; Barberio, Andrea. "Make Your System Firmware Faster, More Flexible and

[6] Kubacki, Michael. "Minimum Platform: Open Source UEFI Firmware for Intel Based Platforms".

Questions?

Lowell Wofford lowell@lanl.gov

Join the Open Source Firmware Slack: https://slack.osfw.dev/

Devon Bautista dbautista@newmexicoconsortium.org