Reliable Linux Kernel Crash Dump with Micro-Controller Assistance

Vasant Hegde
Linux Developer, IBM
hegdevasant@linux.vnet.ibm.com
@hegdevasant

LINUX.CONF.AU
21-25 January
2019
Christchurch, NZ

Outline

- Background
- Self Boot Engine (Micro-Controller)
- SBE assisted dump flow
- Demo
- Advantages

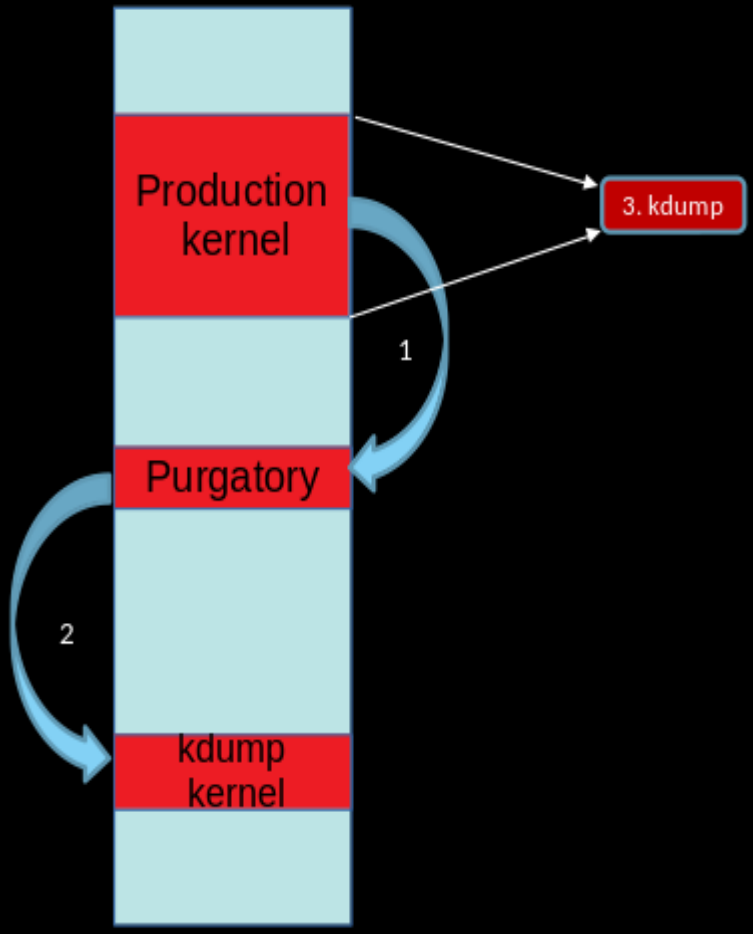
What is dump?

"dump" - snapshot of memory and cpu register state at the time of crash

- Useful for debugging kernel/firmware issues
- No need to recreate issues
- Tools are available to analyze the memory and register contents to root cause reason for the crash

Kernel dumping mechanism : kdump

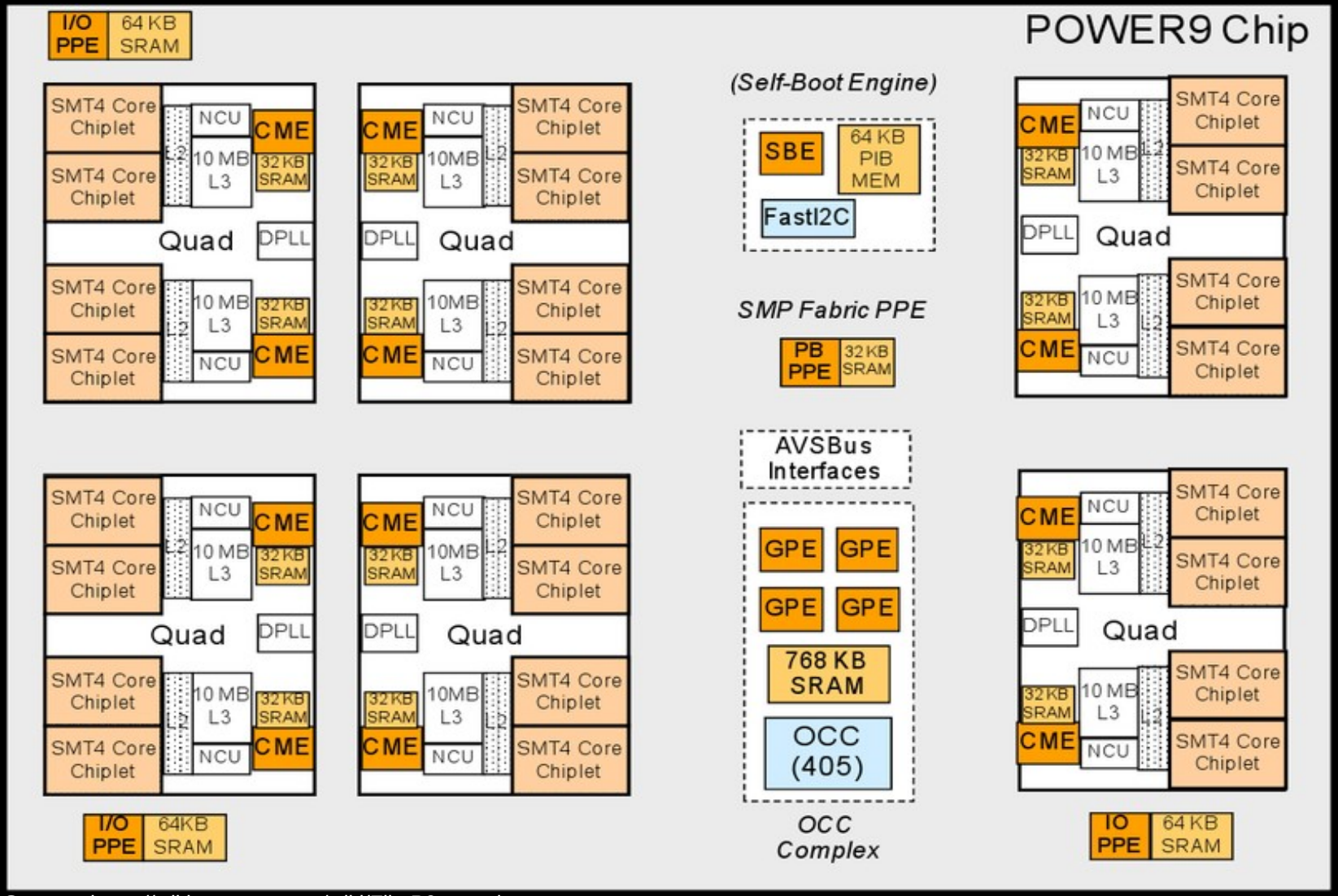
- Linux uses *kdump* as the standard First Failure Data Capture (FFDC)
  - Relies on reserving a portion of memory to boot a small footprint kernel
  - *kdump* is susceptible to flakiness due to
    - Device state inconsistencies
    - Device driver robustness
    - DMAs in flight
    - Buggy software can stomp on reserved memory



Kernel dumping mechanism : fadump

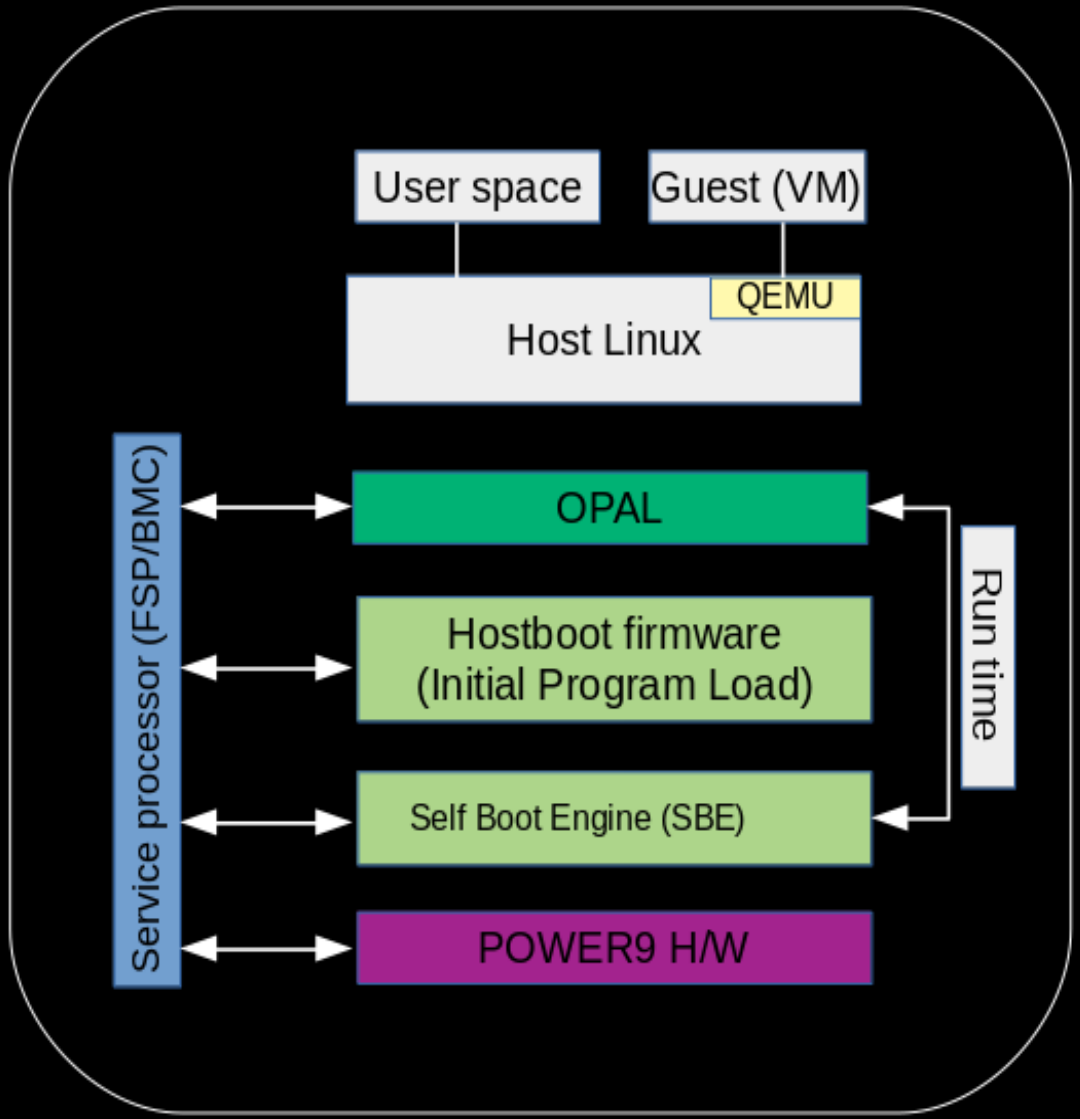
- Power unique Firmware Assisted Dump (fadump) works only on Linux running on IBM PowerVM hypervisor
  - Hypervisor saves state of crashed guest
  - Not applicable to bare metal Linux (OPAL based) systems

POWER9 chip with SBE (Micro-Controller)



Source : https://wiki.raptorcs.com/wiki/File:P9_ppe_instances.png

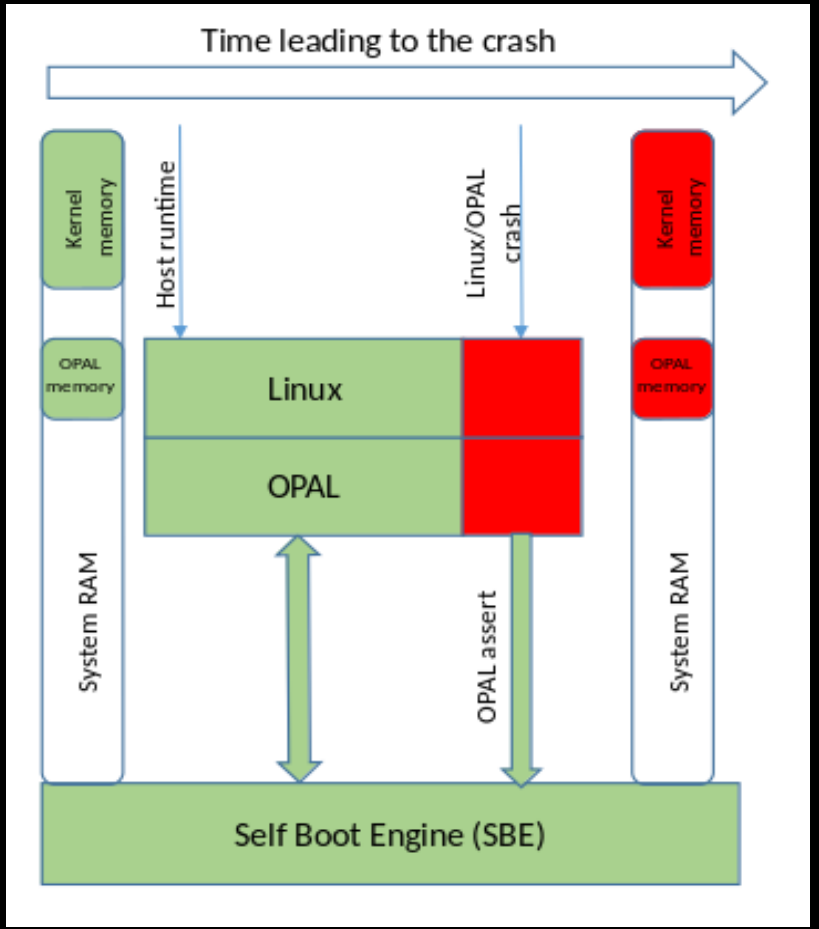
OpenPower System



SBE assisted kernel dump flow – stage 1

"reboot" but preserve memory content and stash CPU registers at the time of crash

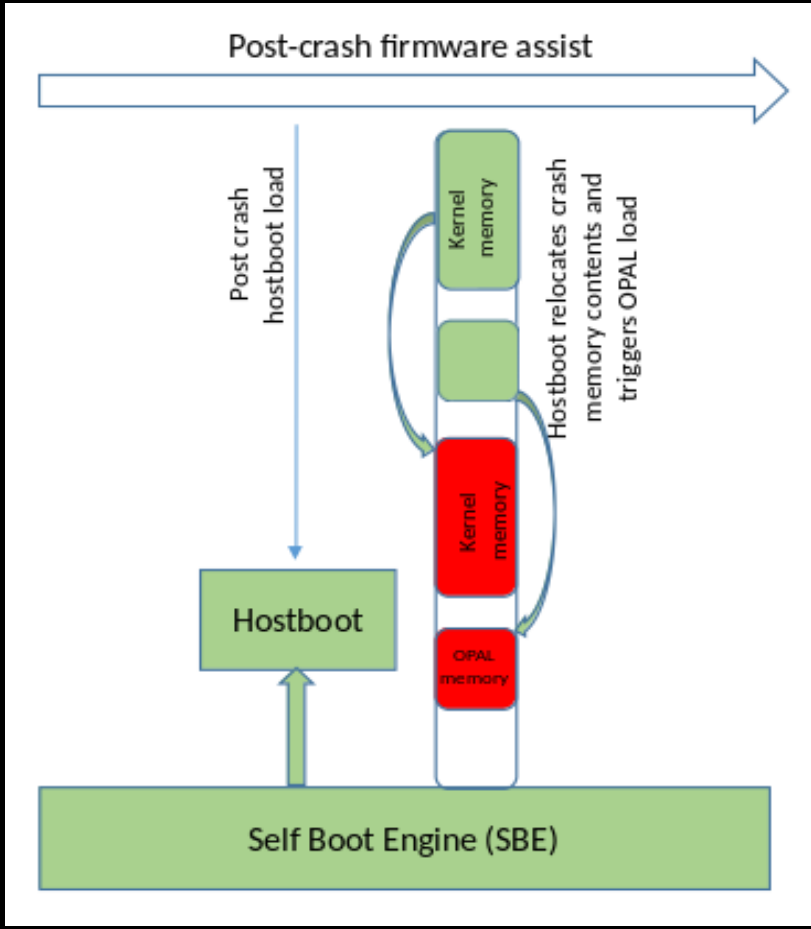
- Flow
  1. OPAL/kernel reserves memory required to capture dump
  2. Kernel registers for SBE assisted dump
  3. Trigger dump (Kernel → OPAL → SBE interrupt)



SBE assisted kernel dump flow – stage 2

"reboot" but preserve memory content and stash CPU registers at the time of crash

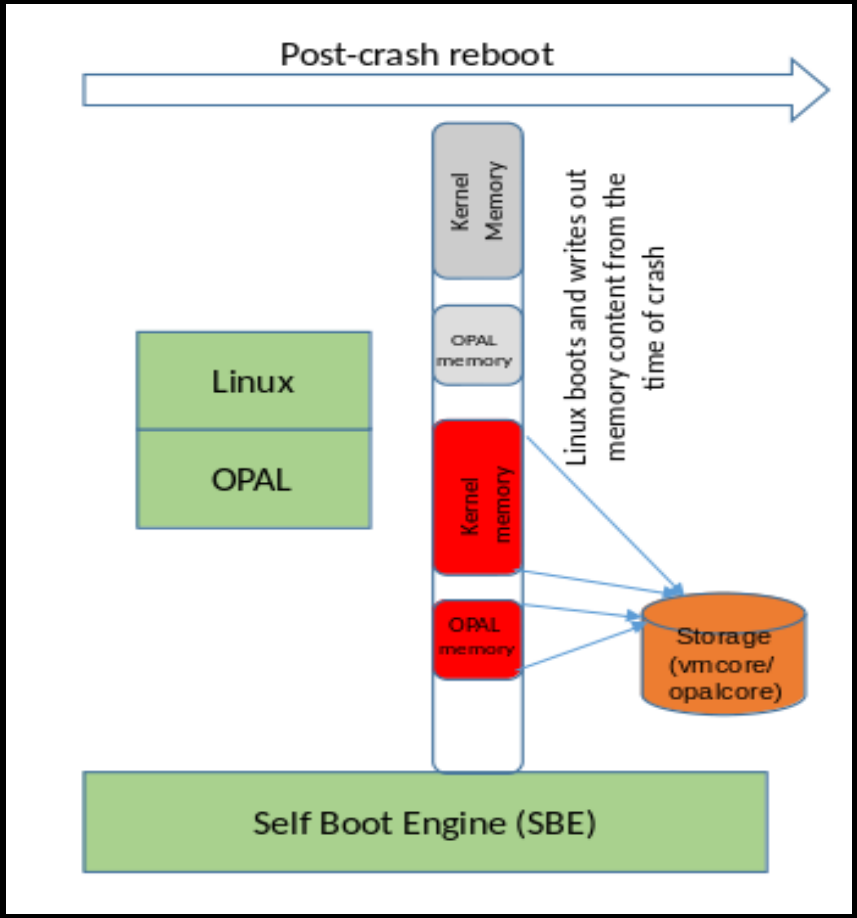
- Flow
  1. SBE salvages relevant CPU register data and exports to OPAL
  2. Hostboot moves OPAL/kernel memory to reserved memory area



SBE assisted kernel dump flow – stage 3

"reboot" but preserve memory content and stash CPU registers at the time of crash

- Flow
  1. OPAL exports saved CPU registers and memory dump to Host Linux
  2. Kernel/kdump copies *vmcore* and *opalcore* to disk
  3. Use *crash* tool to analyze vmcore and *gdb* to analyze opalcore



Demo

Advantages

- First of a Kind (FOAK) in the Linux world where the OS **and** firmware dumps are captured on the **first** instance of a crash
- Firmware assistance allows for OPAL and Linux to load at their default addresses – low possibility of anomalies due to relocation
- Has been made to work with existing Linux kernel dump capture and analysis tools (crash/gdb)
  - Lower maintenance cost
  - No end-user reskilling needed

Current state

- Patches are posted to upstream mailing list
  - OPAL : https://lists.ozlabs.org/pipermail/skiboot/2018-November/012715.html
  - Kernel : https://lists.ozlabs.org/pipermail/linuxppc-dev/2018-December/183187.html

Team

- Amit Tendolkar
- Ananth Narayan
- Dean Sanner
- Daniel Crowell
- Hari Krishna Bathini
- Mahesh Salgaonkar
- Nagendra Gurram
- Raja Das
- Sachin Gupta
- Sampa Misra
- Stewart Smith
- Sunil Kumar
- Vasant Hegde
- Venkatesh Sainath

… and many other IBMers who helped us throughout the development process

Legal Statement

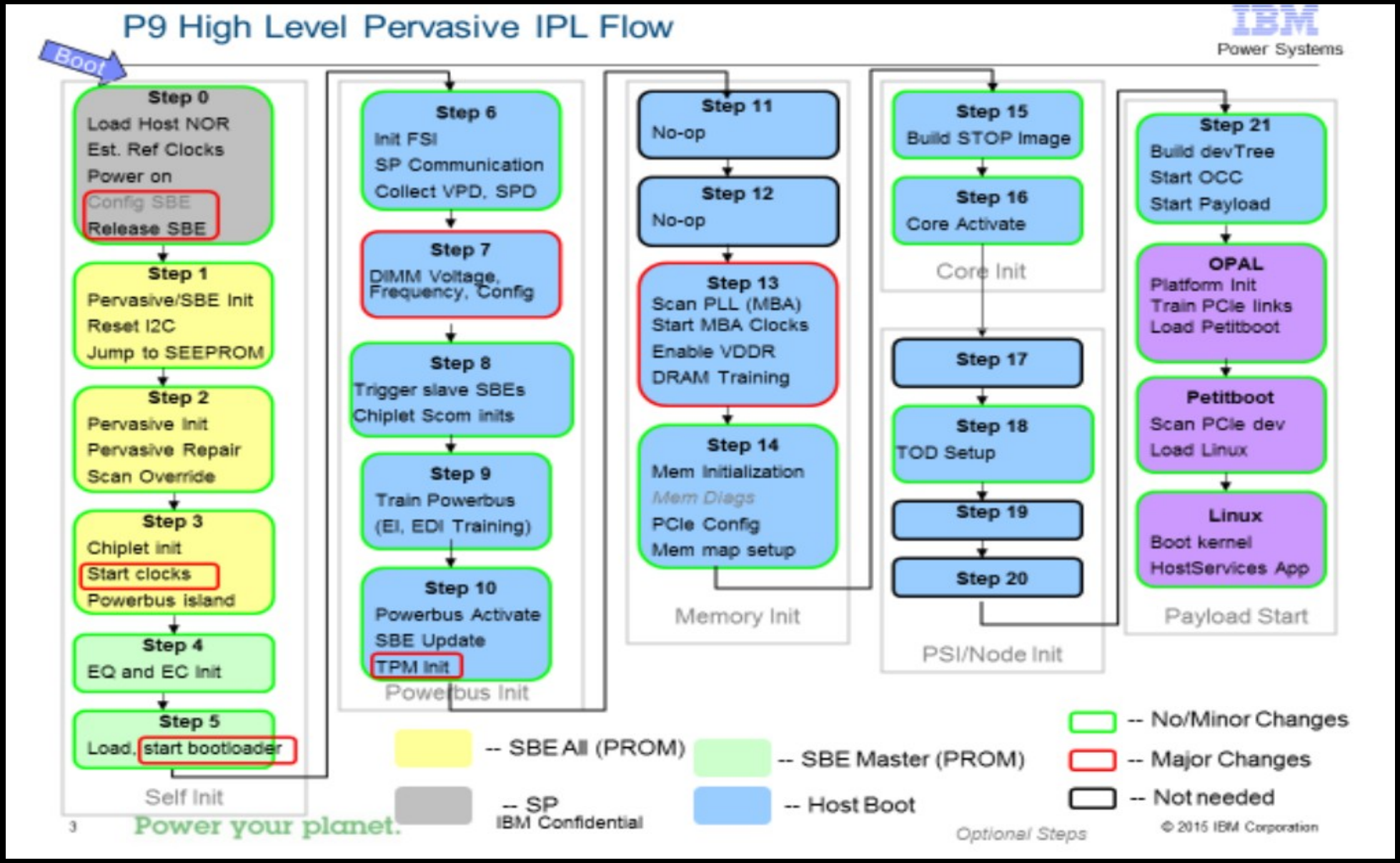
- This work represents the view of the authors and does not necessarily represent the view of their employers.
- IBM and IBM (logo) are trademarks or registered trademarks of International Business Machines Corporation in the United States and/or other countries.
- Linux is a registered trademark of Linus Torvalds.
- Other company, product, and service names may be trademarks or service marks of others.

Questions?

Thank You!

Backup

POWER9 Boot flow



Source: https://github.com/open-power/docs/blob/master/hostboot/P9_Boot_Flow_OpenPOWER.pdf