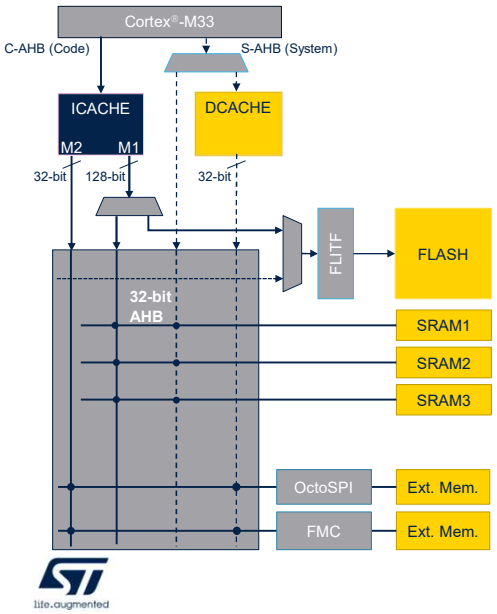
life.augmented

STM32H5

ICACHE

Hello, and welcome to this presentation of the ICACHE module which is embedded in all products of the STM32H5 microcontroller family.

Overview



ICACHE is an 8-KB instruction cache, connected to the C-AHB Code bus of the Cortex®-M33, that improves performance when fetching instructions (or reading data) from internal or external memories

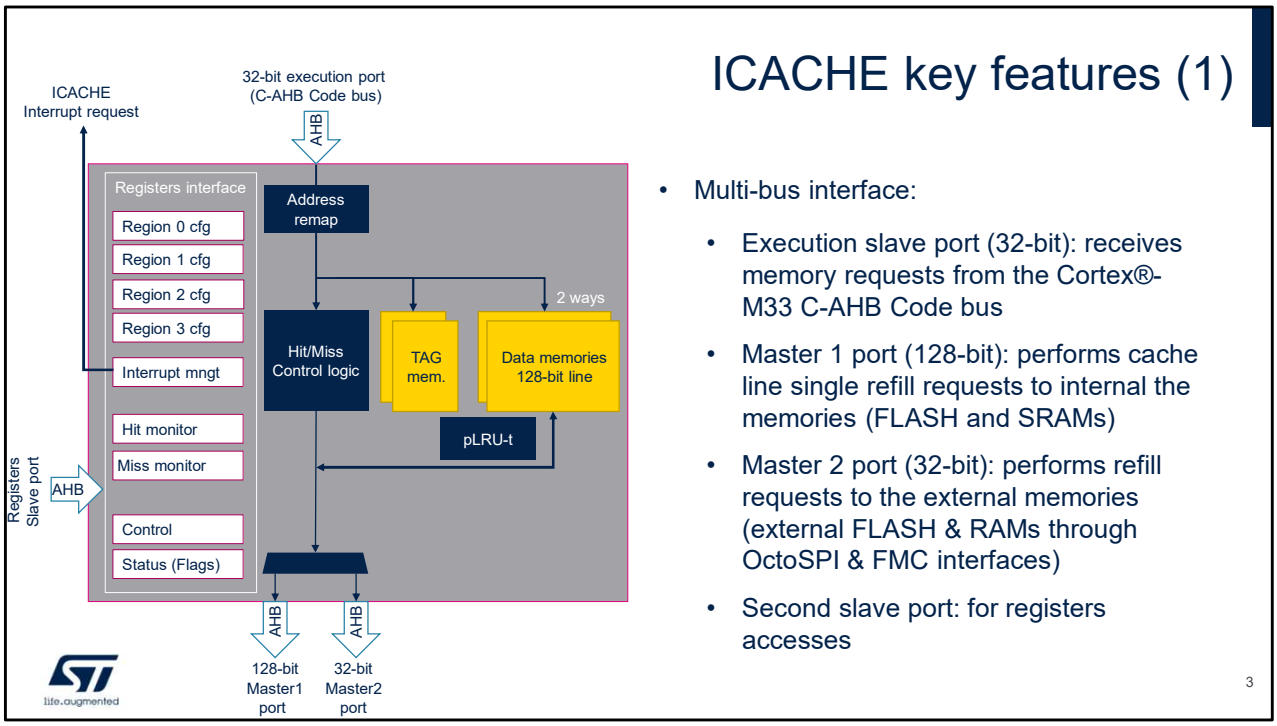
Application benefits

- Higher Performance achieved by close to zero wait-state program fetches
- Remapping logic allows any internal or external memory range to be cached
- Lower power consumption: hitting program fetches from small internal ICACHE, rather than from bigger internal or external memories

2

The instruction cache (ICACHE) is introduced on the C-AHB code bus of the Cortex®-M33 processor to improve performance when fetching instructions and data from internal Flash or SRAM memories or from external memories through the OctoSPI or FMC interfaces.
ICACHE allows a close to zero wait-state performance on program fetches in most use cases, due to intrinsic caching operation.
This performance is achieved through the following two features: hit-under-miss support and critical-word-first refill policy.
The internal flash is accessed by a dedicated 128-bit AHB fast bus (this is the same in STM32U5 microcontrollers, and the only difference compared to the ICACHE

implementation in the STM32L5 microcontrollers).
SRAM 1, 2 and 3, OctoSPI and FMC are accessed through a 32-bit AHB slow interconnect.
Caching internal SRAMs is not recommended, because the cache does not improve the latency.
This two-master architecture decouples the cache refill path from external memories from the high-bandwidth path to the flash memory.
The remapping logic allows four internal or external memory ranges to be cached, by defining for them an alias address in the code section range [0x0000_0000:0x1FFF_FFFF].
The instruction cache reduces the consumption of the microcontroller by accessing instructions and data in the internal ICACHE, rather than from the larger, more power consuming main memories.
Configuring ICACHE as direct-mapped by software allows an even lower power consumption, compared to the 2-way set associative organization, which is also supported.

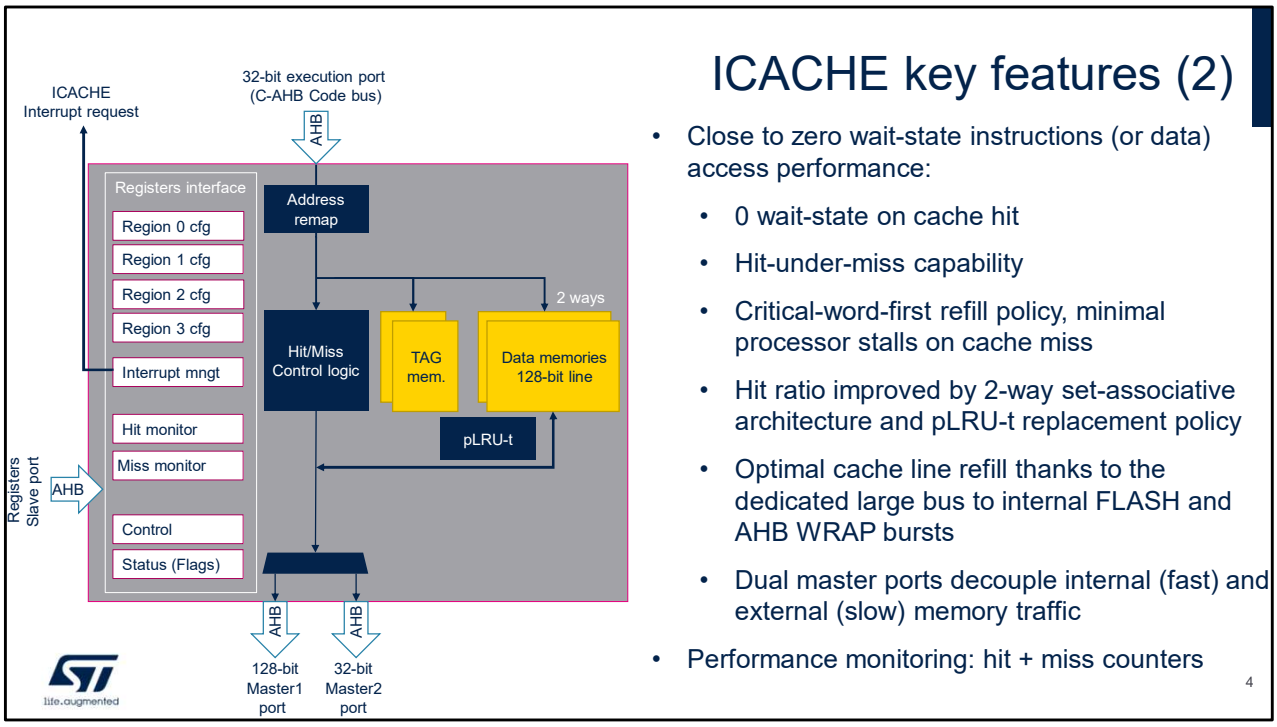


The multi-bus interface minimizes potential conflicts between memory traffic:

- The 32-bit execution slave port receives memory requests from the Cortex®-M33 C-AHB Code bus
- The 128-bit master 1 port performs cache line refills from the internal memories (FLASH and SRAMs)
- The 32-bit master 2 port performs cache line refills from the external memories: external FLASH and RAMs accessed through the OctoSPI and FMC interfaces
- The second slave port is used for registers accesses.

When an external memory access is marked as non-cacheable by the MPU, the ICACHE is bypassed. The request is forwarded to the external memory on the ICACHE master 1 or 2 port in the same clock cycle. Only

the address may be modified due to the address remapping feature.



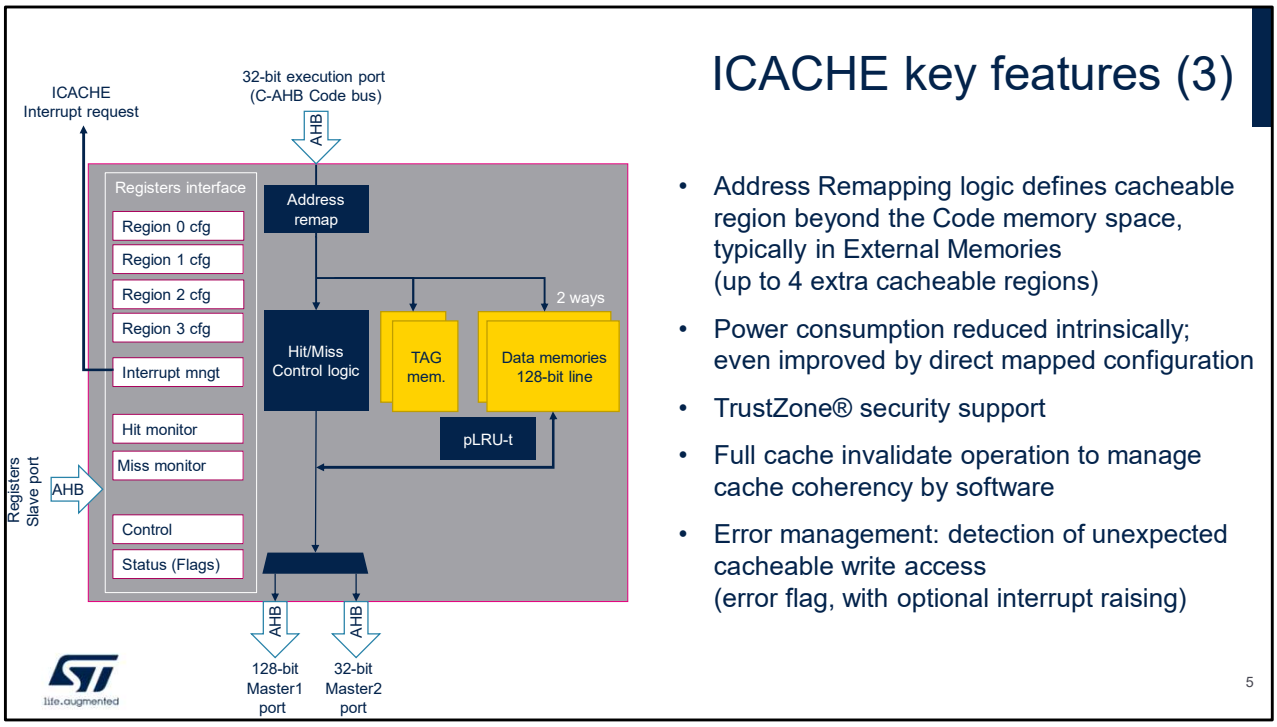
The ICACHE offers close to zero wait states data read/write access performance due to:

- Zero wait-state on cache hit,
- Hit-under-miss capability, that serves new processor requests while a line refill (due to a previous cache miss) is still going on,
- And critical-word-first refill policy, which minimizes processor stalls on cache miss.

The hit ratio is improved by:

- The 2-way set-associative architecture and
- The pseudo-least-recently-used, based on binary tree (or pLRU-t) replacement policy. This algorithm is a good tradeoff between hardware complexity and performance.

Thanks to the wide 128-bit bus, a cacheline refill from flash only requires a single data transfer; because 128 bits represent exactly one 16-byte cacheline.
Cache lines read from the external memories are transferred with the critical word first, by implementing WRAP4 AHB transaction ordering, in order to deliver the instruction requested by the processor's fetcher first.
The dual master port architecture decouples internal and external memory traffic. For example, SRAM fetches are not stalled by cacheline refills from external memories.
Interrupt latency is minimized when the interrupt service routines are located in the internal flash or SRAMs.
The ICACHE implements performance counters: one 32-bit hit counter and one 16-bit miss counter.
This performance monitoring analyzes and optimizes code placement in accordance with cacheability to achieve the most performant code traffic.



The remapping logic is very convenient to extend the cacheable region beyond the 512 Megabyte code memory address range, which starts at address zero.
Up to four external regions can be defined and for each of them the refill port can be selected: either master 1 or master 2.
Coherency is needed when programming the Secure Attribution Unit (SAU) and the Memory Protection Unit (MPU) attributes for both the external regions and their aliased code subregions.
Power consumption is reduced when ICACHE is used: most instruction accesses are performed from internal cache memory rather from main memories.
Configuring the ICACHE as a direct-mapped cache rather

than the default 2-way set associative mode, also contributes to reduce the consumption, because only one cut of tag and data memory is accessed instead of two. However, the direct-mapped organization may affect the performance, when the distance between two programs needed at the same time is an integer multiple of the cache size.
A dedicated Secure-bit in TAG RAM of each cache line prevents non secure requests from hitting secure ICACHE entries.
A invalidate maintenance operation is supported to invalidate the entire contents of the instruction cache, typically when the main memory content is modified.
This operation is controlled by software by accessing a memory-mapped register.
This is a fast command, non interruptible, with an end of operation raising a specific flag and possibly an interrupt.
An error flag and possibly an interrupt are raised whenever an unexpected cacheable write access is received on the execution port.
The ICACHE does not manage AHB bus errors returned to master 1 or master 2 ports. It simply forwards the AHB response received on the master port back to the processor.

SUMMARY

Cache line size	16 bytes
Cache size	8 KB
Organization	2-way set associative Or direct-mapped
Maintenance operations	Invalidate
Number of regions to remap	4
Range granularity of memory regions to be remapped	2 MB

life.augmented

6

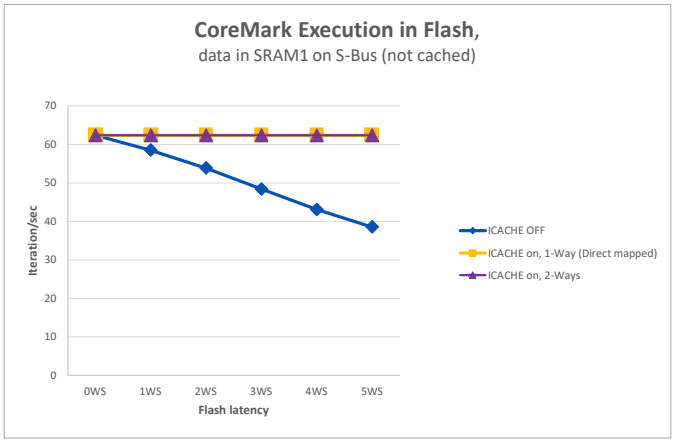
This table summarizes the characteristics of the instruction cache:

- 16 byte cache line size, transferred through a burst transaction of four words or a single data transaction of one quadword
- 2-way set associative, 8-KB cache that can be configured as a direct-mapped cache.

A global invalidate maintenance operation is supported.
ICACHE defines an alias address in the Code region for up to four external memory regions.
The address remapping is applied on the Code alias address, transforming it into the external physical destination address. The minimum region size is 2 megabytes, the maximum size 128 megabytes.

Performance: coremark benchmark

- STM32H5xx tests conditions (results extrapolated from STMU5xx ones):
  - System clock is 16 MHz
  - Code executed in Flash



life.augmented

7

In this chart, the performance in direct-mapped and 2-way set associative modes is the same.
The reason is that the entire benchmark fits into the ICACHE.
Once the code is within the ICACHE, the flash latency has no impact on the performance.
When the ICACHE is disabled, the larger the flash latency, the lower the performance.

ICACHE errors and interrupts

Interrupt vector	Interrupt event	Event Flag	Interrupt Enable bit	Interrupt Clear bit	Description
ICACHE	Functional Error	ICACHE_SR [ERRF]	ICACHE_IER [ERRIE]	ICACHE_FCR [CERRF]	Unsupported cacheable write request detected
	End of Busy State	ICACHE_SR [BSYENDF	ICACHE_IER [BSYENDIE]	ICACHE_FCR [CBSYENDF]	When the cache-busy state is finished, at end of a cache (full) invalidation operation

- ICACHE does not manage AHB bus errors on Master 1 or Master 2 ports transactions, but propagates them back to the Execution port (that received the initial Core C-bus transaction)

life.augmented

8

The two sources of ICACHE global interrupt are:

- Error detection on cacheable write requests, which sets the ERRF bit in the ICACHE status register
- End of the full Invalidate operation, which sets the BSYENDF bit in the ICACHE status register.

There is no ICACHE management of errors occurring on a Master 1 or Master 2 port request.

The erroneous response is propagated through ICACHE back to the Cortex-M33.

Low-power modes

ICACHE clocked on the Cortex®-M33 C-AHB bus clock.

- Same clock domain as the Cortex®-M33 core: same clock frequency and same behavior regarding the power modes

Mode	Description
Run	Active
Sleep	Active
Stop	Frozen, ICACHE register contents are kept ➢ Option: a dedicated control bit in Power Controller to power-down ICACHE (code content lost) in Stop mode
Standby	Powered-down ➢ The peripheral must be reinitialized after exiting Standby mode

When disabled, ICACHE is bypassed, and internal TAG and Data memories are not accessed:

- Almost no power consumption in ICACHE, with the drawback that each instruction is fetched from the more power consuming main memory

life.augmented

ICACHE is clocked at the same frequency as the Cortex M33 core, because the ICACHE only caches instructions requested by the Cortex-M33.
Consequently, the ICACHE and the Cortex-M33 have the same state in the various low power modes.
When the microcontroller is in stop mode, the user can decide to power-down the ICACHE.
When the ICACHE is disabled, the ICACHE is bypassed, except the remapping mechanism that remains functional.
C-AHB bus requests, whether they are remapped or not, are just forwarded to the master ports.
So, the ICACHE consumes less, because TAG and Data memories are not accessed, but each instruction is fetched from the more power consuming targeted main memory.
To reduce power consumption, the performance monitor is disabled by default.

Thank you

life.augmented

In addition to this presentation, you can refer to the following presentations:

- Data cache
- Flash
- FMC
- OCTOSPI.