**Bộ nhớ trên chip:**

* Phần logic lập trình được (PL) cần bộ nhớ trên chip đầu vào, đầu ra vì có thể xảy ra tranh chấp tài nguyên, đồng bộ hóa và có thể truy xuất dữ liệu từ bên ngoài (DDRAM).
* Giảm thời gian truy xuất và có thể dùng luôn bộ nhớ trước đó để sử dụng tính toán.
* DDRAM có dữ liệu lớn, muốn vận chuyển đến lõi tính toán chính thì cần đi qua rất nhiều AXI Bus, DRAM. Từ đó, chúng ta có thể hểu là “Bộ nhớ trên chip đầu vào, đầu ra” như là kho chứa, có thể lấy sẵn và không cần đi qua quá nhiều bước.
* Tránh việc gọi DRAM quá nhiều, thời gian của cta không phụ thuộc vào thời gian di chuyển nữa. => cân bằng pipeline.
* Lõi tính toán chính:
  + Bộ nhớ trên chip cục bộ cần thiết vì có thể tiết kiệm tài nguyên và lưu trữ trạng thái để có thể dùng sau này.
* Giống như tiến trình đợi process.
* Tinh gọn bộ máy.

LUTRAM (Look-up Table) trong FPGA.

* Thường lưu trữ dữ liệu nhỏ, thường là vài chục đơn vị.
* BRAM thường là IP có sẵn.
* Các LUT nằm trong tài nguyên logic CLB (Configurable Logic Books) có thể tạo ra bộ nhớ nhỏ và độ trễ thấp.

BRAM (Block RAM):

* Có trạng thái NO\_CHANGE (không có gì thay đổi), READ\_FIRST (ghi vào bao nhiêu thì hiển thị bấy nhiêu), WRITE\_FIRST (ghi và đọc đồng thời vào cùng một địa chỉ bộ nhớ, phép ghi sẽ được ưu tiên).

FIFO quản lý thứ tự dữ liệu và có hạn chế việc truy xuất dữ liệu không theo thứ tự và bộ nhớ ít, chủ yếu là lưu trữ tạm thời trong thời gian ngắn.

Main Flash để lưu dữ liệu phần mềm hệ thống hoặc hệ điều hành.