

الف) HLT حالتی است که در آن پردازنده کاری انجام نمی دهد (دری وقت را منتظر می دهد)  
 ب) HLT دستور دارد که هرگاه OS آن اجرا کند، وارد این حالت می شود و هرگاه وقتی ای باید از این حالت بیرون می آید.

ج) اگر OS کاری برای انجام دادن نداشته باشد، به این حالت می رود مثلاً وقتی که منتظر دستور I/O است برای کاهش مصرف منابع می تواند به این حالت برود.

الف) باید این دو روش را ترکیب کنیم و هر چند تا I/O را در دست بگیریم و اگر برای هر دو می وقفه اتفاق افتاد، تمامی افعالی کرده را می بینیم.

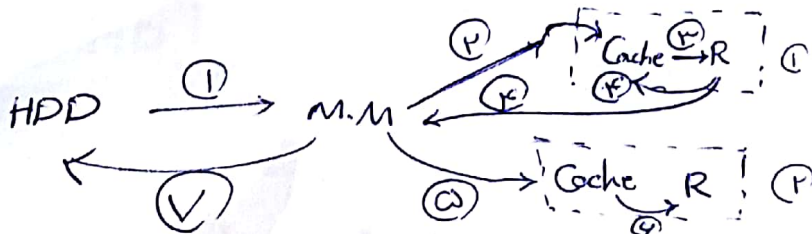
ب) سرکشی برای وقتی که وقفه ها مهم باشند بهتر است برداری برای وقتی که فاصله زمانی زیاد است و مهم نیستند بهتر است.

و روش ما برای تعداد I/O زیاد بهتر است.

نصف افزاری: وقفه هایی که منتظر دریافت ورودی از ماوس یا کیبورد هستیم.  
 نرم افزاری: وقفه هایی که در اجرای یک برنامه دچار خطا می شویم. مثلاً در ساده ترین حالت تقسیم بر صفر.

الف) باید به وقفه با اولویت ها مختلف به هم به صوری که اگر وقفه برای کاربر ۱ است اولویت آن بالاتر از انجام کارهای مربوط به کارهای کاربر ۲ و ۳ می باشد، همچنین اولویت برای کاربر شماره ۲ نیز است ۳ نیز باشد.  
 ب) وقتی که وقفه برای کاربر با اولویت شتر به صورت پشته قبل از انجام کار کاربر با اولویت کمتر برسد و باعث شود کار کاربر با اولویت کمتر همواره ناتمام بماند.

ج) چون این سیستم دارای ۲ پردازنده است، اولین حافظه مشترک بین آنها مموری است پس باید سریع تر داده هم روی گشتاری مموری ایدیت کنیم و بعداً آن را به حافظه شتر ببریم. یک بیت dirty هم باید برای A در نظر بگیریم و با تفسیر مقدار آن به پردازنده دوم بفهمانیم که اگر زودتر برای دریافت مقدار آن آمده است باید صبر کنند تا مقدار آن ایدیت شود.



④ اگر کادهای مربوط به I/O مستقیم باشند به جای اینکه داده اول به CPU برود و پس از آن به MM برود، از DMA استفاده می‌شود که یعنی داده مستقیماً به MM می‌رود. در این بین نیز CPU چند کلاک به صورت قوی و فوری کار انجام می‌دهد تا حالت فضا کردن هم بدست نهد و می‌تواند به Cache خود کار کند که چون DMA کار خودش را می‌کند و CPU هم کار خودش را می‌کند پس باعث افزایش هم‌رندی می‌شود.

⑤ در UMA همه پردازنده‌ها به صورت یکسان به حافظه‌ای با پس یکسان دسترسی دارند و این مسئله باعث این می‌شود که پردازنده‌ها منتظر هم بمانند و سرعت هر کدام و کارایی آن کم شود.

در NUMA هر پردازنده برای خودش یک حافظه دارد و آن حافظه محلی آن پردازنده است که به سرعت بکارت دسترسی دارد ولی هر پردازنده به حافظه محلی پردازنده دیگر دسترسی با سرعت بالایی ندارد و این مسئله باعث می‌شود که به طور کلی سرعت سیستم پایین بیاید ولی با این حال سرعت آن از UMA سریع‌تر است.

①

$$\frac{10 \text{ MB/s}}{4 \text{ KB}} = 2500 \quad \text{تعداد درخواست‌ها به DMA}$$

$$2500 \times 5000 = 1250000 \quad \text{کلاک برای انجام کار}$$

$$\frac{1250000 \text{ کلاک}}{500 \text{ MHz}} = 2.5 \text{ /}$$