INTERDEPARTMENTAL

MASSACHUSETTS INSTITUTE OF TECHNOLOGY *CAMBRIDGE, MASS. 02139*

11/25/66

from the office of

TO: F. J. Corbató
J. H. Saltzer

FROM: Robert R. Fenichel

This memo describes the motivational background of the Load Adjuster. None of it is new.

You may consider my Load Adjuster Process Process to be blocked.

1. General Warning

What happens immediately following a loss of some proper signal fraction of system capacity?

2. On absentee jobs

A. They must be ordered. That is, there is a notion of priority among them. This is so because at any given time there will be more logged in than running, and the system will not simply be commutating among these.

B. First-come-first-served ordering is not enough. A job holding reservations must probably be preferred, and other priorities (e.g., person submitting) must be allowed.

C. Reservation vs no reservation is not enough. In the case of a partial loss of system capacity, execution of some reservation-holding jobs may have to be suspended. Which ones? There must be a complete linear ordering of absentee jobs.

D. Upon the appearance of important work, some absentee jobs may be suspended. This state of suspension, unlike the state of being logged out, should be invisible to the absentee-job-owner.

E. Perhaps the demons are just very important absentee jobs.

Can there be equivalent priority jobs?

3. On interactive jobs

A. There is a notion of machine saturation. Before saturation, no problems are visible. The rest of this section assumes saturation.

B. The interactive jobs must be ordered. There are two reasons here; as with absentee jobs, the first is related to the problem of partial crash: when some users must go, which are chosen?

C. Also, suppose I want to log in. Still assuming saturation, the need for an ordering is evident. I will be able to log in if and only if I have more importance than the least important current user. Whoever he is, he must then be logged out.

D. The same considerations of reservations apply here as with absentee jobs.

4. On avoiding adhocities

A. Are interactive jobs always more important than absentee jobs? Are they always less important? I think neither suggestion is correct.

B. This suggests the existence of a single ordering of all logged-in process-groups. At times of supersaturation through system crash or new logins, process groups at the bottom are suspended or logged out.

C. As presently seen, the Load Adjuster is the means for keeping this ordering going. The L-A consists of

(1) The process-group ranker, which listens to loginabsentee and log-out, says yes/no to login, and trims from the low end in response to

(2) The Autopilot, which decides what trimming is needed in view of the capacity/load ratio reported by

(3) The Load-Detector, which may maintain software dipsticks around the system, or which may just talk to privileged users to get pronouncements.