PACORA: Performance-Aware Convex Optimization for Resource Allocation

Sarah L. Bird and Burton J. Smith

Resource Allocation
Given a set of applications, how many of each resource should the OS give to each application?

Goal: Guarantee Quality-of-Service to applications while maximizing efficiency i.e., Good user experience maximizing battery life

Resources: processing elements, cache slices, memory pages, bandwidth to memory, etc.

PACORA Framework
Construction of the resource allocation problem as a convex optimization

Two functions represent each application
penalty function, runtime function

A user-level scheduler is responsible for managing threads on the resources

The convex optimization tries to minimize the total penalty of the system

Penalty Functions
Represent the importance of an app to the system as a function of the app’s performance

Penalty functions are set by the system to represent the relative cost of missing a deadline for an app
Application deadlines are represented by the service requirement e.g., frame time, responsiveness deadline

Runtime Functions
Convex approximation of the measured performance functions

Performance Model of the Applications
Runtime\(_{(s_0,i)} \rightarrow (s_1,i)\)

(m,i): Allocation of resource of type 1 to App P

App Runtime Model
\( T(w,b,a,m) = \sum_{i} \frac{w_i}{b_i - a_i(m_i)} \)

- \(w\): quantities of work (learned)
- \(b\): allocations of bandwidth resources
- \(a\): bandwidth amplification functions (learned)
- \(m\): allocations of cache or memory resources

Managing Power and Energy
Application 0 can be used to represent the idle resources in the system
Assume all idle resources are powered off

Energy can be managed using Penalty\(_0\)\( (\pi_0)\) and Runtime\(_0\)\( (\tau_0)\)
\( \pi_0\) is defined to be the total system power \( \tau_0\) has a slope that depends on the battery charge

As battery depletes the OS may choose to increase the slope of \( \pi_0\) to reflect the increased value of saving energy

Advantages
Enables the OS to efficiently use resources and provide QoS to applications

- Convex optimization is relatively inexpensive optimization problem with a single extreme point
- Fast, incremental solutions are feasible
- Penalty Function Slopes allow the system to express relative priorities of application
- Priorities change as a function of performance (criticality)
- Penalty Function Intercept encapsulates QoS requirements (responsiveness)
- And additional process can be used to represent system energy