XIV_Approximate analytical models for networked servers subject to MMPP Arrival Process

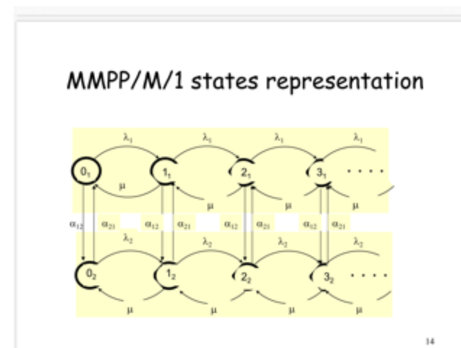
SLA & penalty minimization

Process flow: shows the process in analysis

SLA risk analysis:
1) Def of params in the SLA
2) Workload characterization and svc time ident.
3) platform and resource allocation policy modeling and evaluation
4) Economical risk identification

Markov Modulated Poisson Process

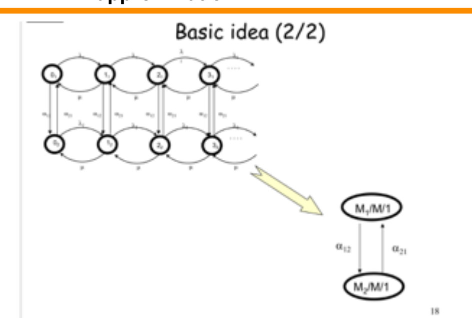
Goal: a technique to evaluate MMPP/M/1 with complexity as M/M/1



Basic idea: model as a combo of M/M/1 process

Approx must be used to evaluate platforms subject to SLA constraints; response time must be less of T (e.g. 3 seconds) for a given P (e.g.95%)

$F_{approximation}(T)<F_{MMP/M/1}(T)<P$



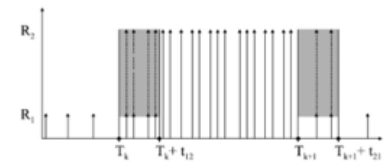
Approximation construction

Identification of H(M1,...,MH)MMPP states
Id of arrival rates
Ev of steady state prob for each state Si of MMPP
Ev of cumulative dist. function
(linear combination of H states Mi/M/1)

Unbiased approximation (response time)

It is approximated using as weights, probabilities pi, for MMPP to stay in each state Si

Unbiased approximation (response time)



$$R = \sum_{i=1}^{H} w_i R_i \qquad w_i = \frac{p_i \lambda_i}{\sum_{j=1}^{H} p_j \lambda_j}$$

Lower bound approximation

Systematic overestimation of the queue lenght during transient periods
Ev of steady state prob for each state Si of MMPP
(using standard results in queuing theory)
Ev of the transient phases durations (classic queuing theory)
Ev of modified probability
Generation of lower bound by performing
a weighted superposition of Mi/M/1

$$p'_i = p_i + \sum_j^{\lambda_i > \lambda_j} p_i \alpha_{ij} t_{ij} - \sum_j^{\lambda_i < \lambda_j} p_j \alpha_{ji} t_j$$

Upper bound approximation

Systematic underestimation of queue ...