to the GI/PH/1 queue. *AAP*, 21:159–180, 1989.
[20] N. B. Shah, K. Lee, and K. Ramchandran. When do redundant requests reduce latency? In *Allerton*, 2013.
[21] B. Snyder. Server virtualization has stalled, despite the hype. http://www.infoworld.com/print/146901, 2010.
[22] A. Vulimiri, P. Godfrey, R. Mittal, J. Sherry, S. Ratnasamy, and S. Shenker. Low latency via redundancy. In *CoNEXT*, 2013.
[23] M. Wu, X.-H. Sun, and H. Jin. Performance under failures of high-end computing. In *ACM/IEEE SC*, page 48, 2007.
[24] Z. Zheng and Z. Lan. Reliability-aware scalability models for high performance computing. In *IEEE CLUSTER*, 2009.

APPENDIX

Proof of Proposition 1

Since jobs only fail during service, the failure probability is given by $p_S = \int_0^\infty \boldsymbol{s}_{\text{ser}} \exp(S_{\text{ser}}x)\boldsymbol{S}_S^* dx = -\boldsymbol{s}_{\text{ser}} S_{\text{ser}}^{-1}\boldsymbol{S}_S^*$. Thus, the probability that a job service time lasts for at most x time units and succeeds is

$$\begin{aligned} F_s(x) &= \frac{1}{p_S}\int_0^x \boldsymbol{s}_{\text{ser}} \exp(S_{\text{ser}}y)\boldsymbol{S}_S^* dy \\ &= 1 + \frac{1}{p_S}\boldsymbol{s}_{\text{ser}} S_{\text{ser}}^{-1} \exp(S_{\text{ser}}x)\boldsymbol{S}_S^*. \end{aligned}$$

This matrix-exponential representation of the service times for successful jobs can be turned into a PH representation by defining a diagonal matrix Π such that $\Pi\mathbf{1}=\eta'$, where $\eta=-\boldsymbol{s}_{\text{ser}}S_{\text{ser}}^{-1}$ is the stationary distribution of the service phase. Defining $B_{\text{ser}}=\Pi^{-1}S'_{\text{ser}}\Pi$, we have

$$\begin{aligned} 1 - F_s(x) &= -\frac{1}{p_S}\boldsymbol{s}_{\text{ser}} S_{\text{ser}}^{-1}\Pi^{-1}\Pi \exp(S_{\text{ser}}x)\Pi^{-1}\Pi\boldsymbol{S}_S^* \\ &= \frac{1}{p_S}\eta\Pi^{-1} \exp(B'_{\text{ser}}x)\Pi\boldsymbol{S}_S^* \\ &= \frac{1}{p_S}\boldsymbol{S}_S^{*\prime}\Pi \exp(B_{\text{ser}}x)\mathbf{1}. \end{aligned}$$

This defines a proper PH distribution as the vector $\boldsymbol{S}_S^{*\prime}\Pi/p_S$ is stochastic. This results from $\boldsymbol{S}_S^*$ and Π being non-negative, and $\boldsymbol{S}_S^{*\prime}\Pi\mathbf{1} = -\boldsymbol{S}_S^{*\prime}(\boldsymbol{s}_{\text{ser}}S_{\text{ser}}^{-1})' = p_S$.

Proof of Theorem 1

The proof of this result follows similar arguments as that of [18, Theorem 1], so we focus on the main differences. As in [18, Theorem 1], we rely on the fact that the PH representation of the waiting time distribution is obtained from a time-reversal argument [19], thus we also use this argument to build the PH representation of the response times. Since we focus on the *successful* jobs only, we follow Proposition 1 to obtain the PH representation of the service time distribution of successful jobs. Further, we split this representation for jobs that wait and jobs that do not. The initial service phase of jobs that wait is given by $\boldsymbol{\alpha}_{\text{busy}} = c\boldsymbol{\pi}_{\text{busy}}(T-S^{(\text{MAP})})$, as this is the distribution of the phase *just after a downward jump in* $X(t)$, and c is a normalizing constant such that $c^{-1}=\boldsymbol{\pi}_{\text{busy}}(T-S^{(\text{MAP})})\mathbf{1}$. Thus we apply a time-reversal by defining the diagonal matrix Δ_{busy} such that $\Delta_{\text{busy}}\mathbf{1}=-\gamma(\boldsymbol{\alpha}_{\text{busy}}S_{\text{ser}}^{-1})'$. We then follow similar steps as in the proof of Proposition1 to obtain the PH representation for jobs that wait $(\boldsymbol{\beta}_{\text{ser}}^{\text{busy}}, B_{\text{ser}}^{\text{busy}})$ as

$$\boldsymbol{\beta}_{\text{ser}}^{\text{busy}}=\boldsymbol{S}_S^*\Delta_{\text{busy}}/p_S, \quad B_{\text{ser}}^{\text{busy}}=\Delta_{\text{busy}}^{-1}S'_{\text{ser}}\Delta_{\text{busy}}.$$

A similar result is obtained for jobs that do not wait, considering that in this case jobs start service according to $(1-\gamma)\boldsymbol{\pi}(0)$. We thus define Δ_{idle} as a diagonal matrix such that $\Delta_{\text{idle}}\mathbf{1}=-(1-\gamma)(\boldsymbol{\pi}(0)S_{\text{ser}}^{-1})'$, and the corresponding PH representation $(\boldsymbol{\beta}_{\text{ser}}^{\text{idle}}, B_{\text{ser}}^{\text{idle}})$ is given by

$$\boldsymbol{\beta}_{\text{ser}}^{\text{idle}}=\boldsymbol{S}_S^*\Delta_{\text{idle}}/p_S, \quad B_{\text{ser}}^{\text{idle}}=\Delta_{\text{idle}}^{-1}S'_{\text{ser}}\Delta_{\text{idle}}.$$

With these PH representations for the service time, we obtain Eq. (5) by putting together the paths of jobs that do not wait, which start service with $\boldsymbol{\beta}_{\text{ser}}^{\text{idle}}$, with those of jobs that wait, which start service with $\boldsymbol{\beta}_{\text{ser}}^{\text{busy}}$. Given the time-reversal, the response time of jobs that wait is composed of a first stage of service followed by a second stage of waiting. Further, the phase in which the service stage ends determines the stage in which the waiting stage begins. The remaining of the proof follows the same steps as that of [18, Theorem 1], such that the matrix $\tilde{P}_{s,w}=\Gamma^{-1}(T-S^{(\text{MAP})})'\Lambda$ is a stochastic matrix that determines how the phase at the end of the service phase determines the phase at the beginning of the waiting phase. Once a job starts the waiting phase, it evolves according to S_{wait} until absorption. Further details can be found in [18].