We start with the classical definition of parallel efficiency:

$$E=\frac{T_1}{T_1+T_o} \quad ,$$

where T_1 is the execution (processing) time of a volume of computational work spent by 1 CPU and measured in seconds, T_o is the execution (processing) overhead time needed to complete this volume of work by p CPUs and measured in seconds.

By definition, this time scales are related to T_p (the execution / processing time needed to complete this volume of work by p CPUs) via the relation:

$$pT_p = T_1 + T_o \quad .$$

And if we denote a volume of computational work as M (measured in bits) and processing throughput of a node as G (bits/s), then $T=M/G$. By supposing a situation, where the same amount of work M/p is distributed among different processing nodes that complete this work within interval T_k , we get:

$$pT_p = \sum_k T_k = \frac{M}{p} \sum G_{p,k}^{-1} \quad ,$$

and efficiency E can be calculated from the measurements of $G_{p,k}$ and G_1 as:

$$E = \frac{pG_1^{-1}}{\sum_k G_{p,k}^{-1}} \quad .$$