In this contribution, a new framework for H2-optimal reduction is presented, motivated by the local nature of both (tangential) interpolation and H2-optimal approximations. The main advantage is given by a decoupling of the cost of reduction from the cost of optimization, resulting in a significant speedup in H2-optimal reduction. In addition, a middle-sized surrogate model is produced at no additional cost and can be used e.g. for error estimation. Numerical examples illustrate the new framework, showing its effectiveness in producing H2-optimal reduced models at a far lower cost than conventional algorithms. Detailed discussions and optimality proofs are presented for applying this framework to the reduction of multiple-input, multiple-output linear dynamical systems. The paper ends with a brief discussion on how this framework could be extended to other system classes, thus indicating how this truly is a general framework for interpolatory H2 reduction.